# Posterior asymptotics of nonparametric location-scale mixtures for multivariate density estimation

ANTONIO CANALE[*] and PIERPAOLO DE BLASI[**]

*Department of Economics and Statistics, University of Turin and Collegio Carlo Alberto, Italy.*
*E-mail: [*]antonio.canale@unito.it; [**]pierpaolo.deblasi@unito.it*

Density estimation represents one of the most successful applications of Bayesian nonparametrics. In particular, Dirichlet process mixtures of normals are the gold standard for density estimation and their asymptotic properties have been studied extensively, especially in the univariate case. However, a gap between practitioners and the current theoretical literature is present. So far, posterior asymptotic results in the multivariate case are available only for location mixtures of Gaussian kernels with independent prior on the common covariance matrix, while in practice as well as from a conceptual point of view a location-scale mixture is often preferable. In this paper, we address posterior consistency for such general mixture models by adapting a convergence rate result which combines the usual low-entropy, high-mass sieve approach with a suitable summability condition. Specifically, we establish consistency for Dirichlet process mixtures of Gaussian kernels with various prior specifications on the covariance matrix. Posterior convergence rates are also discussed.

*Keywords:* Bayesian nonparametrics; density estimation; Dirichlet mixture; factor model; posterior asymptotics; sparse random eigenmatrices

## 1. Introduction

Multivariate density estimation is a fundamental problem in nonparametric inference being also the starting point for nonparametric regression, clustering, and robust estimation. For modeling continuous densities, standard nonparametric Bayes methods rely on Dirichlet process (DP) [10] mixture models of the form

$$f(x) = \int K(x; \theta) \, dP(\theta), \qquad P \sim \mathrm{DP}(\alpha P^*), \tag{1}$$

where $K(x; \theta)$ is a probability kernel depending on some finite-dimensional parameter $\theta$ and $\mathrm{DP}(\alpha P^*)$ is a Dirichlet process with total mass $\alpha > 0$ and $P^*$ a probability measure over the space of parameters $\theta$. Model (1) has been introduced by Lo [21] and made popular by Escobar and West [9], Müller, Erkanli and West [25]. For densities on $\mathbb{R}^d$, a common choice for the kernel $K(x; \theta)$ is the normal density $\phi_\Sigma(x - \mu)$, that is

$$\phi_\Sigma(x - \mu) = (2\pi)^{-d/2} \det(\Sigma)^{-1/2} \exp\left\{ -\tfrac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu) \right\},$$

for $\mu \in \mathbb{R}^d$ a $d$-dimensional vector of locations and $\Sigma$ a positive definite symmetric matrix of variance-covariances. Depending on whether or not the mixture involves the scale parameter $\Sigma$, we will speak of location and location-scale mixtures, respectively. In the former case, all components in the mixture share the same $\Sigma$ which is typically modeled with an independent prior.

In the univariate case, the asymptotic properties of DP mixtures are well known for both location [11,13,20,29] and location-scale mixtures [14,28]. In the multivariate case, the only results, to our knowledge, are confined to the case of location mixtures. Posterior consistency is studied in [32] assuming a truncated inverse-Wishart prior for $\Sigma$. Shen, Tokdar, and Ghosal [27] improve considerably on these results by deriving adaptive posterior convergence rates and remove the artificial truncation of the Wishart prior, which prevents an effective implementation. One of the main tools for achieving their result is the use of the stick-breaking representation of the DP to build a low-entropy, high-mass sieve on the space of mixed densities, a procedure whose extension to the multivariate case is a challenging task.

The lack of asymptotic results for mutivariate location-scale mixtures is somehow in contrast with their predominant use in applications [4,17,22,25]. In this paper, we fill this gap in the current literature by establishing posterior consistency for DP location-scale mixtures of multivariate normals with ready to verify conditions on the prior parameter $P^*$. In particular, a distinctive condition with respect to the location mixtures case involves the existence of moments to a certain order of the ratio between the largest and smallest eigenvalues of $\Sigma$, a quantity known in random matrix literature as condition number. This moment condition is satisfied by the inverse-Wishart as well as by other prior specifications that enable scaling to higher dimension. The consistency result exploits the sieve construction suggested by [27] by adapting a convergence rate theorem of [13]. Such an adaptation allows to relax the growth condition on the entropy of the sieve through a summability condition which suitably weighs entropy numbers with prior probabilities, an idea first appeared in [20] and [29]. We also discuss some of the technical issues related to the challenging task of deriving posterior convergence rates for heavy tailed densities together with some preliminary results.

The layout of the paper is as follows. In Section 2, sufficient conditions, on the true $f_0$ and on the prior, to obtain posterior consistency of DP location-scale mixtures are given. In Section 3, some particular prior specifications satisfying such conditions are discussed. Section 4 is about convergence rates and the paper ends with a final discussion.

## 2. Posterior consistency

For any $d \times d$ matrix $A$ with real eigenvalues, let $\lambda_1(A) \geq \cdots \geq \lambda_d(A)$ denote its eigenvalues in decreasing order and $\|A\|_2 = \max_{x \neq 0} \|Ax\|/\|x\|$ be its spectral norm. We denote by $\mathcal{S}$ be the space of $d \times d$ positive definite matrices and by $\mathcal{P}$ the space of probability measures on $\mathbb{R}^d \times \mathcal{S}$. We consider DP location-scale mixtures of the type

$$f_P(x) = \int \phi_\Sigma(x - \mu) \, dP(\mu, \Sigma), \qquad P \sim \mathrm{DP}(\alpha P^*), \tag{2}$$

where $P^* \in \mathscr{P}$ with $\mu$ and $\Sigma$ independent under $P^*$. In marginalizing out $P$, one can write model (2) as

$$f_P(x) = \sum_{h=1}^{\infty} \pi_h \phi_{\Sigma_h}(x - \mu_h), \qquad (\mu_h, \Sigma_h) \overset{\text{i.i.d.}}{\sim} P^*, \qquad \pi_h = V_h \prod_{k<h}(1 - V_k), \qquad (3)$$

where $V_h \overset{\text{i.i.d.}}{\sim} \text{beta}(1, \alpha)$. We denote by $\Pi^*$ the DP prior on $\mathscr{P}$ and by $\Pi$ the prior induced by (2) on the space $\mathscr{F}$ of density functions on $\mathbb{R}^d$.

As customary in Bayesian asymptotics, we take the data $X_1, \ldots, X_n$ to be i.i.d. from some "true" density $f_0 \in \mathscr{F}$ and study the behavior of the posterior as $n \to \infty$ with respect to the $n$-products measure $F_0^n$, $F_0$ being the probability measure associated to $f_0$. As metrics on $\mathscr{F}$ we consider the Hellinger $d(f, g) = \{\int(\sqrt{f} - \sqrt{g})^2\}^{1/2}$, and the $L_1$ $\|f - g\|_1 = \int|f - g|$, which induce equivalent topologies in view of $d^2(f, g) \leq \|f - g\|_1 \leq 2d(f, g)$. By posterior consistency at $f_0$ we mean that, for any $\varepsilon > 0$,

$$\Pi\big(\{f : \rho(f_0, f) > \varepsilon\}|X_1, \ldots, X_n\big) \to 0$$

in $F_0^n$-probability where $\rho$ is either the Hellinger or the $L_1$-metric.

It is known that, for the posterior distribution to accumulate around $f_0$, one has to establish first some support condition of the prior. We say that $\Pi$ satisfies the Kullback–Leibler (KL) property at $f_0$ if

$$\Pi\left\{f : \int \log(f_0/f) f_0 \leq \eta\right\} \geq 0, \qquad \text{for any } \eta > 0. \qquad (4)$$

The KL property has been established in Theorem 5 of [31] for location-scale mixtures of Gaussian kernels with scalar covariance matrices, i.e. $\Sigma = \sigma^2 I$ for $\sigma^2 > 0$ and $I$ the $d \times d$ identity matrix. Minor adaptations are needed to extend this result to the case of non scalar $\Sigma$. Sufficient conditions for (4) involve a mild requirement on the weak support of the DP prior $\Pi^*$ on $\mathscr{P}$ together with some regularity assumptions on $f_0$ like the existence of moments up to a certain order. As for the weak support of $\Pi^*$, we shall assume that the prior mean $P^*$ of the DP is supported on all $\mathbb{R}^d \times \mathcal{S}$, where we take as distance the sum of the Euclidean norm on $\mathbb{R}^d$ and the spectral norm on $\mathcal{S}$. All priors considered in Section 3 satisfy this requirement. As for $f_0$, the same regularity conditions of [31], Theorem 5, apply; they are repeated, for readers' convenience, in Lemma 1 whose proof is reported in the Appendix.

**Lemma 1.** *Let $f_0 \in \mathscr{F}$ and $\Pi$ denote the prior on $\mathscr{F}$ induced by (2). Assume that $f_0$ satisfies the following conditions: $0 < f_0(x) < M$ for some constant $M$ and all $x \in \mathbb{R}^d$; $|\int f_0(x) \log f_0(x)\, dx| < \infty$; for some $\delta > 0$, $\int f_0(x) \log \frac{f_0(x)}{\phi_\delta(x)}\, dx < \infty$, where $\phi_\delta(x) = \inf_{\|t-x\|<\delta} f_0(t)$; for some $\eta > 0$, $\int \|x\|^{2(1+\eta)} f_0(x)\, dx < \infty$. Then $\Pi$ satisfies (4).*

The KL property (4) plays a very important role in consistency since it provides a lower bound for the denominator of the posterior probability. However, posterior consistency in non-compact spaces, such as $\mathscr{F}$, requires an additional condition on the prior which involves the metric entropy of $\mathscr{F}$ (see below for a formal definition). A critical step is to introduce a compact subset $\mathscr{F}_n$,

called sieve, which is indexed by the sample size $n$ and eventually grows to fill the entire parameter space as $n \to \infty$. According to Theorem 2 of [11], the metric entropy of $\mathcal{F}_n$ has to grow slower than linearly in $n$, while the prior probability assigned to $\mathcal{F}_n^c$, the complement of the sieve, needs to decrease exponentially fast in $n$. See Theorem 2.1 of [12] for similar ideas applied to posterior convergence rates. The choice of the sieve is a delicate issue in multivariate density estimation, since the metric entropy tends to blow up with the dimension $d$. A novel sieve construction has been introduced in [27] and it has proven successful in deriving adaptive posterior convergence rates in the case of location mixtures with independent inverse-Wishart prior on $\Sigma$. In particular, the sieve relies on the stick-breaking representation of the DP as in (3). In adapting this sieve construction to location-scale mixtures (see Lemma 2 in the Appendix), the tail behavior of $P^*$ with respect to the condition number, that is, the ratio of the maximum and minimum eigenvalue of $\Sigma$, plays a crucial role. A straight application of Theorem 2 of [11] would require a too restrictive condition on this tail behavior, ruling out common choices for the part of $P^*$ involving $\Sigma$ like the inverse-Wishart distribution. See Remark 2 in the Appendix for a detailed explanation. For this reason, we resort to a different posterior consistency theorem which consists in a modification of Theorem 5 of [13]. The main idea is to relax the growth condition on the entropy of $\mathcal{F}_n$ through a summability condition of entropy numbers weighted by square roots of prior probabilities. This modification can be applied also to relax the usual exponential tail behavior of the marginal of $P^*$ on the location parameters $\mu$, with a weaker power tail decay. As pointed out by [13], there is a trade off between entropy and summability which is worth exploring in Bayesian asymptotics. Our result is a step in this direction. Similar ideas had earlier appeared in [20] and in [29].

To state the posterior convergence result, we recall the definition of entropy of $\mathcal{G} \subset \mathcal{F}$ as $\log N(\varepsilon, \mathcal{G}, d)$ where $N(\varepsilon, \mathcal{G}, d)$ is the minimum integer $N$ for which there exists $f_1, \ldots, f_N \in \mathcal{F}$ such that $\mathcal{G} \subset \bigcup_{j=1}^{N} \{f : d(f, f_j) < \varepsilon\}$.

**Theorem 1.** *Suppose* $\mathcal{F}_n \subset \mathcal{F}$ *can be partitioned as* $\bigcup_j \mathcal{F}_{n,j}$ *such that, for* $\varepsilon > 0$,

$$\Pi\big(\mathcal{F}_n^c\big) \lesssim \mathrm{e}^{-bn}, \qquad \text{for some } b > 0, \tag{5}$$

$$\sum_j \sqrt{N(2\varepsilon, \mathcal{F}_{n,j}, d)} \sqrt{\Pi(\mathcal{F}_{n,j})} \mathrm{e}^{-(4-c)n\varepsilon^2} \to 0, \qquad \text{for some } c > 0. \tag{6}$$

*Then* $\Pi(f : d(f_0, f) > 8\varepsilon | X_1, \ldots, X_n) \to 0$ *in* $F_0^n$*-probability for any* $f_0$ *satisfying* (4).

The proof is similar to that of Theorem 5 of [13] and is presented in the Appendix. Here we state and prove the main result on posterior consistency for DP location-scale mixtures of Gaussian kernels.

**Theorem 2.** *Let* $f_0$ *satisfy the conditions of Lemma 1. Consider the prior* $\Pi$ *defined in* (2) *with* $P^*$ *that satisfies the following tail behaviors: for some positive constants* $c_1, c_2, c_3, r > (d-1)/2$ *and* $\kappa > d(d-1)$,

$$P^*\big(\|\mu\| > x\big) \lesssim x^{-2(r+1)}, \tag{7}$$

$$P^*\big(\lambda_1\big(\Sigma^{-1}\big) > x\big) \lesssim \exp\big(-c_1 x^{c_2}\big), \tag{8}$$

$$P^*\big(\lambda_d\big(\Sigma^{-1}\big) < 1/x\big) \lesssim x^{-c_3},\tag{9}$$

$$P^*\big(\lambda_1\big(\Sigma^{-1}\big)/\lambda_d\big(\Sigma^{-1}\big) > x\big) \lesssim x^{-\kappa}\tag{10}$$

*for all sufficiently large $x > 0$. Then the posterior is consistent at $f_0$.*

Condition (7) is weaker than the usual exponential tail condition in [28] (univariate case) and in [27] (location mixture case). The distinctive condition for multivariate location-scale mixtures turns out to be (10) as explained in Remark 2. It corresponds to the existence of the $d(d-1)+1$ moment of the quantity $\lambda_1(\Sigma^{-1})/\lambda_d(\Sigma^{-1})$ which is known in random matrix theory as condition number of $\Sigma^{-1}$. See [8] for a recent contribution on the topic. In particular, the condition number plays an important role in electronical engineering problems in the context of multiple-input multiple-output communication systems. See [23] for related results.

**Proof of Theorem 2.** The proof is an application of Theorem 1 and is based on the entropy upper bounds of Lemma 2 in the Appendix. Let $M_n = \underline{\sigma}_n^{-2c_2} = n$ and $H_n = \lfloor Cn\varepsilon^2/\log n \rfloor$ for a positive constant $C$ to be determined later. Also, let $\mathbf{j} = (j_1, \ldots, j_{H_n})$, $j_h \in \mathbb{N}^*$, and $\mathbf{l} = (l_1, \ldots, l_{H_n})$, $l_h \in \mathbb{N}$. Define

$$\mathcal{F}_n = \left\{ f_P \text{ with } P = \sum_{h \geq 1} \pi_h \delta_{(\mu_h, \Sigma_h)} : \sum_{h > H_n} \pi_h \leq \varepsilon; \right.$$

$$\left. \underline{\sigma}_n^2 \leq \lambda_d(\Sigma_h), \lambda_1(\Sigma_h) \leq \underline{\sigma}_n^2 (1 + \varepsilon/\sqrt{d})^{M_n}, \text{ for } h \leq H_n \right\},$$

$$\mathcal{F}_{n,\mathbf{j},\mathbf{l}} = \left\{ f_P \in \mathcal{F}_n : \sqrt{n}(j_h - 1) < \|\mu_h\| \leq \sqrt{n} j_h, n^{2^{l_h-1}\mathbb{1}_{(l_h \geq 1)}} < \frac{\lambda_1(\Sigma_h)}{\lambda_d(\Sigma_h)} \leq n^{2^{l_h}}, \text{ for } h \leq H_n \right\}$$

so that $\mathcal{F}_n \uparrow \mathcal{F}$ as $n \to \infty$ and $\mathcal{F}_n \subset \bigcup_{\mathbf{j},\mathbf{l}} \mathcal{F}_{n,\mathbf{j},\mathbf{l}}$.

As for (5) of Theorem 1, note that

$$\Pi\big(\mathcal{F}_n^c\big) \leq \Pr\left\{ \sum_{h > H_n} \pi_h > \varepsilon \right\} + H_n \big[ P^*\big(\lambda_d(\Sigma) < \underline{\sigma}_n^2\big) + P^*\big(\lambda_1(\Sigma) > \underline{\sigma}_n^2 (1 + \varepsilon/\sqrt{d})^{M_n}\big) \big].$$

Use the stick-breaking representation of the DP for the first term (see the proof of Proposition 2 of [27]), (8) and (9) to get

$$\Pi\big(\mathcal{F}_n^c\big) \lesssim \left\{ \frac{e\alpha}{H_n} \log \frac{1}{\varepsilon} \right\}^{H_n} + H_n \left[ e^{-c_1 \underline{\sigma}_n^{-2c_2}} + \underline{\sigma}_n^{-2c_3}\left(1 + \frac{\varepsilon}{\sqrt{d}}\right)^{-c_3 M_n} \right],\tag{11}$$

so that

$$\Pi\big(\mathcal{F}_n^c\big) \lesssim \big(Cn\varepsilon^2/\log n\big)^{-Cn\varepsilon^2/\log n} + \big(Cn\varepsilon^2/\log n\big)\big[ e^{-c_1 n} + n^{c_3/c_2}(1 + \varepsilon/\sqrt{d})^{-c_3 n} \big]$$

$$\lesssim \exp\big\{-\big(Cn\varepsilon^2/\log n\big)\log\big(Cn\varepsilon^2/\log n\big)\big\} + e^{-c_1 n} + \exp\big\{-c_3 n \log(1 + \varepsilon/\sqrt{d})\big\}.$$

Note that $(Cn\varepsilon^2/\log n)\log(Cn\varepsilon^2/\log n) > Cn\varepsilon^2/2$ for large enough $n$, therefore

$$\Pi(\mathcal{F}_n^c) \lesssim e^{-Cn\varepsilon^2/2} + e^{-c_1 n} + e^{-c_3 \log(1+\varepsilon/\sqrt{d})n} \lesssim e^{-bn}$$

for $0 < b < \min\{C\varepsilon^2/2, c_1, c_3 \log(1 + \varepsilon/\sqrt{d})\}$. Hence (5) is satisfied.

We next show that $\mathcal{F}_{n,\mathbf{j},\mathbf{l}}$ satisfies the summability condition (6) of Theorem 1. By an application of Lemma (2) in the Appendix,

$$N\big(\varepsilon, \mathcal{F}_{n,\mathbf{j},\mathbf{l}}, \|\cdot\|_1\big)$$
$$\lesssim \exp\bigg\{ d H_n \log M_n + H_n \log \frac{C_1}{\varepsilon}$$
$$+ \sum_{h \le H_n} \log\bigg[\bigg(\frac{\sqrt{n}\,j_h}{\underline{\sigma}_n \varepsilon/2} + 1\bigg)^d - \bigg(\frac{\sqrt{n}(j_h - 1)}{\underline{\sigma}_n \varepsilon/2} - 1\bigg)^d\bigg] + \frac{d(d-1)}{2}\log \frac{2dn^{2^{l_h}}}{\varepsilon^2}\bigg\}$$

for some positive $C_1$. Let $c_4 = 1/2 + 1/(2c_2)$ and note that

$$\bigg(\frac{\sqrt{n}\,j_h}{\underline{\sigma}_n \varepsilon/2} + 1\bigg)^d - \bigg(\frac{\sqrt{n}(j_h - 1)}{\underline{\sigma}_n \varepsilon/2} - 1\bigg)^d = \bigg(\frac{2n^{c_4}\,j_h}{\varepsilon} + 1\bigg)^d - \bigg(\frac{2n^{c_4}\,j_h}{\varepsilon} + 1 - \frac{2n^{c_4}}{\varepsilon} - 2\bigg)^d$$
$$\lesssim d\bigg(\frac{2n^{c_4}}{\varepsilon} + 2\bigg)\bigg(\frac{2n^{c_4}\,j_h}{\varepsilon} + 1\bigg)^{d-1} \lesssim \frac{n^{c_4 d}\,j_h^{d-1}}{\varepsilon^d},$$

where inequality sign "$\lesssim$" is for both large $n$ and $j_h$. Hence, we have

$$N\big(\varepsilon, \mathcal{F}_{n,\mathbf{j},\mathbf{l}}, \|\cdot\|_1\big) \lesssim \exp\bigg\{ d H_n \log n + H_n \log \frac{C_1}{\varepsilon}$$
$$+ \sum_{h \le H_n} \log \frac{n^{c_4 d}\,j_h^{d-1}}{\varepsilon^d} + \frac{d(d-1)}{2}\log \frac{2dn^{2^{l_h}}}{\varepsilon^2}\bigg\}. \tag{12}$$

Moreover, by using tail conditions (7) and (10),

$$\Pi(\mathcal{F}_{n,\mathbf{j},\mathbf{l}}) \le \prod_{h \le H_n} P^*\big(\|\mu\| > \sqrt{n}(j_h - 1), \lambda_1(\Sigma)/\lambda_d(\Sigma) > n^{2^{l_h-1}\mathbb{1}_{(l_h \ge 1)}}\big)$$
$$\lesssim \prod_{h \le H_n} \big[\sqrt{n}(j_h - 1)\big]^{-\mathbb{1}_{(j_h \ge 2)}2(r+1)} n^{-\mathbb{1}_{(l_h \ge 1)}2^{l_h-1}\kappa} \tag{13}$$

with the convention $0^0 = 1$. A combination of (12), (13) and $d^2(f, g) \le \|f - g\|_1$ imply that $\sqrt{N(2\varepsilon, \mathcal{F}_{n,\mathbf{j},\mathbf{l}}, d)}\sqrt{\Pi(\mathcal{F}_{n,\mathbf{j},\mathbf{l}})}$ is bounded by a multiple of

$$\exp\bigg\{\frac{d + c_4 d}{2}Cn\varepsilon^2\bigg\}$$
$$\times \prod_{h \le H_n} j_h^{(d-1)/2}\big[\sqrt{n}(j_h - 1)\big]^{-\mathbb{1}_{(j_h \ge 2)}(r+1)} n^{(d(d-1)/4)2^{l_h}} n^{-\mathbb{1}_{(l_h \ge 1)}2^{l_h-1}(\kappa/2)}. \tag{14}$$

By hypothesis, $r > (d-1)/2$ so that $K := \sum_{j \geq 2} j^{(d-1)/2}(j-1)^{-(r+1)} < \infty$. Hence by summing (14) with respect to $\mathbf{j}$, we get

$$\exp\left\{\frac{d + c_4 d}{2} C n \varepsilon^2\right\} \left(1 + n^{-(r+1)/2} K\right)^{H_n} \prod_{h \leq H_n} n^{(d(d-1)/4) 2^{l_h}} n^{-\mathbb{1}_{(l_h \geq 1)} 2^{l_h - 1}(\kappa/2)}. \quad (15)$$

Moreover, note that

$$\sum_{l \geq 0} n^{(d(d-1)/4) 2^l} n^{-\mathbb{1}_{(l \geq 1)} 2^{l-1}(\kappa/2)}$$

$$= n^{d(d-1)/4} + \sum_{l \geq 1} \exp\left\{-\frac{d(d-1)}{2} \log n \left(\frac{\kappa}{d(d-1)} - 1\right) 2^{l-1}\right\}$$

$$\leq 2 n^{d(d-1)/4}$$

for $n$ large enough since, by hypothesis, $\kappa/[d(d-1)] - 1 > 0$. Hence, by summing (15) with respect to $\mathbf{l}$ we get the upper bound

$$\exp\left\{\frac{d + c_4 d}{2} C n \varepsilon^2\right\} \left(1 + n^{-(r+1)/2} K\right)^{H_n} \left[2 n^{d(d-1)/4}\right]^{H_n}$$

$$\leq \exp\left\{\frac{d + c_4 d}{2} C n \varepsilon^2\right\} 4^{H_n} \exp\left\{C n \varepsilon^2 \frac{d(d-1)}{4}\right\} \lesssim \exp\left\{\frac{1}{2}\left[d + c_4 d + \frac{d(d-1)}{2}\right] C \varepsilon^2 n\right\},$$

where in the first inequality we have used $n^{-(r+1)/2} K \leq 1$ for $n$ large enough. By taking $C$ sufficiently small to satisfy $C < 2(4-c)/[d + c_4 d + d(d-1)/2]$ for some $c > 0$, (6) is satisfied and the proof is complete. $\qquad\square$

## 3. Illustration

Theorem 2 holds for DP location-scale mixtures of multivariate Gaussian kernels with minimal requirements on the prior parameter $P^*$. The power tail decay (7) for $\mu$ is indeed important as it covers the prior specification of [25]:

$$P^*(\mu, \Sigma) = N(\mu; m, B) \mathrm{IW}(\Sigma; \Sigma_0, \nu),$$

where an additional hyperprior on $B$ is given by $\mathrm{IW}(B; B_0, \nu_B)$, with $\mathrm{IW}(B_0, \nu_B)$ denoting the inverse-Wishart distribution with scale parameter $B_0$ and $\nu_B$ degrees of freedom. In this case, $\mu$ under $P^*$ has multivariate student's $t$-distribution with $\nu_B$ degree of freedom so that $P^*(\|\mu\|^2 > x) \sim x^{-(\nu_B - d + 1)/2}$ for $x$ large enough. Hence, $\nu_B > 2d$ is sufficient for (7) to hold with $r > (d-1)/2$.

In what follows, we focus on the prior specification of the scale parameter $\Sigma$ and show that conditions (8)–(10) are verified for some important choices which make a substantial practical difference in applications, particularly in high-dimensional settings. Henceforth, we refer to the

prior specification for $\Sigma$ as $\Sigma \sim \mathcal{L}$ meaning that $\mathcal{L}$ is the marginal of $\Sigma$ with respect to the prior mean $P^*$. All proofs are reported in the Appendix.

The first result is about $\mathcal{L}$ being the inverse-Wishart distribution $\mathrm{IW}(\Sigma_0, \nu)$. While conditions (8) and (9) are always satisfied, see, for example, the proof of Lemma 1 of [27], $\nu$ needs to be sufficiently large for tail condition (10) on the condition number to hold. Corollary 1 makes this statement precise.

**Corollary 1.** *Assume $f_0$ satisfies the conditions of Lemma* 1. *Consider a prior $\Pi$ induced by* (2) *with $P^*$ satisfying* (7) *and $\mathcal{L} = \mathrm{IW}(\Sigma_0, \nu)$ with $\nu > 2d(d-1)+d-1$. Then conditions* (8)–(10) *are satisfied and the posterior is consistent at $f_0$.*

There is a rich literature providing alternatives to the inverse-Wishart prior when the dimension of the data is large. For example, a commonly used and successful approach consists on analytic factorizations [3,30] where

$$\Sigma = \Gamma\Gamma^T + \Omega, \qquad \Gamma \sim \mathcal{L}_\Gamma, \qquad \Omega \sim \mathcal{L}_\Omega, \tag{16}$$

where $\Gamma$ is a $d \times r$ matrix with $r < d$, independent from the $d \times d$ diagonal matrix $\Omega$. Let $\gamma_{jh}$ be the $(j,h)$th element of $\Gamma$ (factor loading) and $\sigma_j^2$ be the $j$th diagonal element of $\Omega$ (residual variance). The specification of $\mathcal{L}_\Gamma$ and $\mathcal{L}_\Omega$ corresponds then to a distribution for $\gamma_{jh}$ and $\sigma_j^2$. The next corollary addresses the case of normal factor loadings with inverse gamma distributed residual variances. It turns out that conditions (8)–(9) are automatically satisfied, while a constraint on the shape parameter of the inverse gamma prior is needed for the verification of condition (10).

**Corollary 2.** *Assume $f_0$ satisfies the conditions of Lemma* 1. *Consider a prior $\Pi$ induced by* (2) *with $P^*$ satisfying* (7) *and $\mathcal{L}$ induced by* (16). *Assume that $\gamma_{ij} \overset{\text{i.i.d.}}{\sim} N(0,1)$ and $\sigma_j^{-2} \overset{\text{i.i.d.}}{\sim} \mathrm{Ga}(a,b)$ with $a > d(d-1)$. Then conditions* (8)–(10) *are satisfied and the posterior is consistent at $f_0$.*

***Remark 1.*** Motivated by the need of a method that scales for increasing dimension, [2] introduce a sparse Bayesian factor model for the estimation of high-dimensional covariance matrices according to (16). The model consists of a multiplicative gamma process shrinkage prior on the factor loadings, that is,

$$\gamma_{jh}|\phi_{jh}\tau_h \sim N\big(0, \phi_{jh}^{-1}\tau_h^{-1}\big), \qquad \phi_{jh} \sim \mathrm{Ga}(3/2, 3/2), \qquad \tau_h = \prod_{l=1}^{h} \delta_l,$$

$$\delta_1 \sim \mathrm{Ga}(a_1, 1), \qquad \delta_l \sim \mathrm{Ga}(a_2, 1), \qquad l > 1, \qquad \sigma_j^{-2} \sim \mathrm{Ga}(a, b), \qquad j = 1, \ldots, d.$$

Using similar arguments to Corollary 2 it can be proved that conditions (8)–(10) are satisfied for $a > d(d-1)$.

One can easily build $\mathcal{L}$ which satisfy (8)–(10) by modeling directly the distribution of the eigenvalues. The idea is to use the spectral decomposition of $\Sigma$:

$$\Sigma = O\Lambda O^T, \qquad O \sim \mathcal{L}_O, \Lambda \sim \mathcal{L}_\Lambda, \tag{17}$$

where $O$ is a $d \times d$ orthogonal matrix independent from $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with $\lambda_i > 0$. It is clear that the verification of (8)–(10) involves only the distribution $\mathcal{L}_\Lambda$. Motivated by high dimensional sparse random matrices modeling, Cron and West [5] propose a similar approach. The authors set $O_{i,j}$ to be the rotation matrix for the rotator angle $\omega_{i,j}$ and $O = \prod_{i<j} O_{i,j}(\omega_{i,j})$ and define prior on $O$ through a prior on the rotator angles $\omega$ which naturally accommodates sparsity, namely

$$p(\omega) = \beta_{\pi/2}\mathbb{1}(\omega = \pi/2) + (1 - \beta_{\pi/2})\beta_0\mathbb{1}(\omega = 0) + (1 - \beta_{\pi/2})(1 - \beta_0)p_c(\omega),$$

$$p_c(\omega) = c(\kappa)\exp\{\kappa \cos^2 \omega\}\mathbb{1}(|\omega| < \pi/2).$$

This class of models is particularly useful to induce sparsity without the assumption of a reduced dimensional latent factor and hence can be appealing in many practical situations. The next corollary discusses sufficient conditions on $\mathcal{L}_\Lambda$ to obtain posterior consistency under formulation (17).

**Corollary 3.** *Assume $f_0$ satisfies the conditions of Lemma 1. Consider a prior $\Pi$ induced by (2) with $P^*$ satisfying (7) and $\mathcal{L}$ following (17) with $\lambda_i^{-1} \sim \text{Ga}(a, b)$, $a > d(d - 1)$. Then conditions (8)–(10) are satisfied and the posterior is consistent at $f_0$.*

## 4. Posterior convergence rates

In this section, we discuss some relevant issues related to the derivation of posterior convergence rates for location-scale mixtures. Under the prior conditions (7)–(10), the sieve construction laid down in the proof of Theorem 2 would adapt to any rate $\varepsilon_n = n^{-\gamma}$ (for any $\gamma \in (0, 1/2)$ and up to a logarithmic term) determined by the prior concentration rate. This statement is made precise in the following proposition, whose proof is reported in the Appendix.

**Proposition 1.** *Let $\tilde{\varepsilon}_n = n^{-\gamma}(\log n)^t$ for some $\gamma \in (0, 1/2)$ and $t \geq 0$ and suppose that*

$$\Pi\left(f : \int f_0 \log(f_0/f) \leq \tilde{\varepsilon}_n^2, \int f_0 \log^2(f_0/f) \leq \tilde{\varepsilon}_n^2\right) \geq e^{-n\tilde{\varepsilon}_n^2}. \tag{18}$$

*Then, under the hypothesis of Theorem 2, $\Pi_n\{f : d(f_0, f) > (\log n)^s \tilde{\varepsilon}_n\} \to 0$ in $F_0^n$-probability for any $s > 0$.*

As for the prior contraction rate (18), in the case of location mixtures the derivation of adaptive $\tilde{\varepsilon}_n$ for $\beta$-Hölder $f_0$ consists of three main steps: (i) construct a density $h_\sigma$ (depending on $f_0$ and $\beta$) such that $d(f_0, \phi_\sigma \star h_\sigma) \lesssim \sigma^\beta$ as $\sigma \to 0$; (ii) modify $h_\sigma$ to a compactly supported density $\tilde{h}_\sigma$ with the same approximation properties; (iii) approximate the continuous mixture $\phi_\sigma \star \tilde{h}_\sigma$ by a discrete mixture $\phi_\sigma \star F_\sigma$ with a convenient lower bound of the ratio $\phi_\sigma \star F_\sigma/f_0$. Here $\phi_\sigma \star h$ denotes the convolution of $h$ and $\phi_{\sigma^2 I}$. See [27] and [18]. Steps (ii) and (iii) rely on the assumption of exponential tail of $f_0$, which makes these techniques not suitable for location-scale mixtures, since the latter are more flexible than location mixtures in modeling densities with heavy tails. We come back to this point later. Before, we present a result of the type of convergence rates

which are attainable by adapting these techniques to our setting. To this aim, refer to the definition of the locally $\beta$-Hölder class with envelope $L$, denoted $\mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$, from [27]. Moreover, for a multi-index $k = (k_1, \ldots, k_d)$, $k_i \in \mathbb{N}^*$, define $k_. = k_1 + \cdots + k_d$ and let $D^k$ denote the mixed partial derivative operator $\partial^{k_.}/\partial^{k_1} \cdots \partial^{k_d}$. Finally, we make the following two assumptions on the prior mean $P^*$. As for the scale parameter, we resort to an assumption analogous to (4) in [27] about the mass in neighborhood of the eigenvalues of $\Sigma$: there exist $\kappa^*, a_4, a_5, b_4, C_3$ such that for any $s_1 \geq \cdots \geq s_d \geq 0$ and $t \in (0, 1)$,

$$P^*\big(s_i < \lambda_i\big(\Sigma^{-1}\big) < s_i(1+t), i = 1, \ldots, d\big) \gtrsim b_4 s_d^{a_4} t^{a_5} \exp\big\{-C_3 s_1^{\kappa^*/2}\big\}. \tag{19}$$

As for the mean parameter, we assume that the marginal of $P^*$ has density $f^*$ and that $f^*$ has tails no lighter than the Gaussian, that is, for $b^*, \tau^* > 0$ and $\|x\|$ sufficiently large,

$$f^*(x) \gtrsim \mathrm{e}^{-b^*\|x\|^{\tau^*}}. \tag{20}$$

The proof of the following proposition is deferred to the Appendix.

**Proposition 2.** *Let $f_0 \in \mathcal{C}^{\beta,L,\tau_0}(\mathbb{R}^d)$ be a bounded probability density function satisfying $F_0(|D^k f_0|/f_0)^{(2\beta+\varepsilon)/k_.} < \infty$, $k_. \leq \lfloor \beta \rfloor$, $F_0(L/f_0)^{(2\beta+\varepsilon)/\beta} < \infty$ for some $\varepsilon > 0$ and*

$$f_0(x) \leq c \exp\big\{-b\|x\|^\tau\big\}, \tag{21}$$

*for some $b, c > 0$, $\tau \geq 0$ and $\|x\|$ sufficiently large. Assume that conditions (19) and (20) hold for $\tau^* \leq \tau$. Then (18) holds for*

$$\tilde{\varepsilon}_n = n^{-\beta/(2\beta+d+\kappa^*)}(\log n)^t, \qquad t \geq \frac{d(1 + (\kappa^*+1)/\beta + 1/\tau)}{2 + (d+\kappa^*)/\beta}. \tag{22}$$

Proposition 2 yields, via Proposition 1, a posterior convergence rate which is suboptimal with respect to the minimax rate $n^{-\beta/(2\beta+d)}$ by a term which depends on the constant $\kappa^*$ appearing in (19). For illustration, $\kappa^* = 2$ if, under $P^*$, $\Sigma$ has Inverse Wishart distribution $\mathrm{IW}(\Psi, \nu)$ with $\nu$ degrees of freedom and a positive definite scalar matrix $\Psi$, see Lemma 1 in [27]. Another useful specification is to consider that, under $P^*$, each $\lambda_i(\Sigma^{-1})$ have been independently assigned the distribution of the square of an inverse gamma random variable. Then $\kappa^* = 1$, which leads to a better convergence rate, still suboptimal. The technical reason for which the minimax rate is not achieved is to be found in the prior probability of $L_1$-balls around the frequencies of the approximating mixture of the true density. Specifically, the approximating mixture has all covariance matrices equal to $\sigma_n^2 I$ for a scaling factor $\sigma_n$ which goes to zero as a function of $\tilde{\varepsilon}_n$. In location-scale mixtures $P^*$ put mass proportional to $\exp\{-C_3\sigma_n^{-\kappa^*}\}$ in neighborhoods of $\sigma_n^2 I$ and this accounts for an extra factor in the radius of the $L_1$-ball with respect to the location mixture case. This, in turns, determines worse probability estimates of Kullback–Leibler neighborhoods. See the proof of Proposition 2 for details. It can be proved that the minimax rate $n^{-\beta/(2\beta+d)}$ is recovered upon setting the marginal of $P^*$ on $\Sigma$ depending on $n$ by the scaling factor equal to $n^{-1/(2\beta+d)}(\log n)^{-1/\beta}$, however this has clearly a limited relevance since the rate would be non adaptive and it would rule out the Inverse-Wishart distribution as well as any other commonly

used prior specification for the covariance matrix. The suboptimal convergence rate might be related to the fact that location-scale mixtures are more robust to tails than location mixtures. This should emerge through posterior convergence rates which are slower still adaptive to the minimax rate of a suitably defined class of densities with heavy tails. To this aim, an approximation scheme different than the one used for location mixture is in order since the latter relies on the exponential tail condition (21). This remains an open problem and is left as an argument for future research.

We discuss next how minimax rates for density estimation depend on the tail. It is well known that smoothness alone is not sufficient in order to guarantee consistency of density estimators in the $L_1$-norm. In fact, there is a vast literature on minimax and adaptive minimax density estimation with $L_p$ norm which indicates the existence of a tail zone, i.e. a range of values of $p$, for which the minimax rate depends on $p$ and deteriorates to 1 as $p$ decreases to 1. Such phenomenon does not appear in density estimation on a compact domain, see [16] for an up to date literature review. Goldenshluger and Lepski [16] have also determined a tail dominance condition which illustrates how the tail zone shrinks to an empty set according to the tails of the density. The result which is relevant to our study is about the minmax rate under $L_1$-norm for $\beta$-Holder classes of densities which is, up to $\log n$ factors, given by

$$\max\left\{n^{-\beta/(2\beta+d)}, n^{-(1-\theta)/(1+d/\beta)}\right\}, \tag{23}$$

where $\theta$ is a positive parameter in $(0, 1]$ which determines the heaviness of the tails. See Remark 4.3 in [16]. The fastest rate $n^{-\beta/(2\beta+d)}$ is recovered for $\theta < \beta/(2\beta+d)$ (light tail), while for $\theta \geq \beta/(2\beta+d)$ (heavy tail) a wide range of slower minimax rates is obtained. See Theorem 13 of [6] for a closely related result in the light tail case. Clearly the exponential tail assumption (21) corresponds to small value of $\theta$ and this explain why in [27] the usual $n^{-\beta/(2\beta+d)}$ rate is achieved.

## 5. Discussion

In this paper, we have discussed asymptotic properties of DP location-scale mixtures of Gaussian kernels for multivariate density estimation. To our knowledge, this is the first contribution to posterior asymptotics in the context of multivariate location-scale mixtures, a modelling approach which is a common practice in many applications. We have given sufficient conditions on the DP prior measure on the space of means and covariance matrices in order to achieve posterior consistency. While showing that these conditions are satisfied for widely used inverse-Wishart distribution, with the same practical motivation of providing theoretical justification for models used in practice, we showed that the conditions hold if one uses priors that parsimoniously model the covariance in high dimensional settings, such as a factor model or spectral decomposition having both computational tractability and better fit in finite samples. Future work will deal with mixtures of more general random probability measures, like the two-parameter Poisson–Dirichlet process [26], the normalized inverse Gaussian process [19] and, in general, Gibbs-type priors [15]. A characteristic feature of Gibbs-type priors is given by their heavy-tailedness, which has certain advantages in terms of statistical inference. See [7]. From the perspective of frequentist asymptotics, this implies that the sieve based on the truncated stick-breaking construction does not carry over in a straightforward way and hence this problem deserves further investigations.

# Appendix

**Proof of Lemma 1.** It is well known that $P \in \mathcal{P}$ belongs to the weak support of $\Pi^*$ (meaning that any weak neighborhood of $P$ has positive probability under $\Pi^*$) if the support of $P$ is contained in the support of $P^*$. We recall that the support of $P$ is the smallest closed set of $P$-measure 1. Hence, since the support of $P^*$ is the whole space $\mathbb{R}^d \times \mathcal{S}$, the weak support of $\Pi^*$ is the whole space $\mathcal{P}$ and it contains, in particular, any compactly supported $P$.

The proofs of Theorem 5 and Theorem 2 of [31] imply that, under the conditions of Lemma 1, for any $\varepsilon > 0$, there exists $P_\varepsilon \in \mathcal{P}$ such that $\int f_0 \log(f_0/f_{P_\varepsilon}) \leq \varepsilon$. In particular, $P_\varepsilon$ can be taken as compactly supported, that is $P_\varepsilon(D) = 1$ where $D = [-a, a]^p \times \{\Sigma \in \mathcal{S} : \underline{\sigma}^2 \leq \lambda_i(\Sigma) \leq \bar{\sigma}^2, i = 1, \ldots, d\}$ for some constants $a > 0$, $0 < \underline{\sigma} < \bar{\sigma}$. Since

$$\int f_0(x) \log \frac{f_0(x)}{f(x)} \, dx = \int f_0(x) \log \frac{f_0(x)}{f_{P_\varepsilon}(x)} \, dx + \int f_0(x) \log \frac{f_{P_\varepsilon}(x)}{f(x)} \, dx$$

in order to prove (4), we next show that there exists $W \subset \mathcal{P}$ with $\Pi^*(W) > 0$ such that, for any $P \in W$, $\int f_0 \log(f_{P_\varepsilon}/f_P) \leq \varepsilon$. To this aim, we verify the hypotheses of Lemma 3 in [31]. It is clear that $P_\varepsilon$ belongs to the weak support of $\Pi^*$. Next, condition (A7) of Lemma 3 in [31] requires $\log f_{P_\varepsilon}$ and $\log \inf_{(\mu, \Sigma) \in D} \phi_\Sigma(x - \mu)$ to be $f_0$-integrable. Note that, for $\|x\| < a$, $\inf_{(\mu, \Sigma) \in D} \phi_\Sigma(x - \mu)$ is bounded, while, for $\|x\| \geq a$,

$$\inf_{(\mu, \Sigma) \in D} \phi_\Sigma(x - \mu) = \bar{\sigma}^{-d} \exp\left\{ -\frac{4\|x\|^2}{2\underline{\sigma}^2} \right\}.$$

Hence, under the hypotheses made, $\log \inf_{(\mu, \Sigma) \in D} \phi_\Sigma(x - \mu)$ is $f_0$-integrable. Also

$$|\log f_{P_\varepsilon}| \leq \left| \log \left\{ \bar{\sigma}^{-d} \exp\left\{ -\frac{4\|x\|^2}{2\underline{\sigma}^2} \right\} P_\varepsilon(D) \right\} \right|$$

for $\|x\| \geq a$, so that $\log f_{P_\varepsilon}$ is $f_0$-integrable by a similar argument.

As for condition (A8) of Lemma 3 in [31], it is obviously satisfied since the multivariate normal kernel $\phi_\Sigma(x - \mu)$ is bounded away from zero for $x$ in a compact set of $\mathbb{R}^d$ and $(\mu, \Sigma) \in D$. Finally, condition (A9) of Lemma 3 in [31] requires that, for $C \subset \mathbb{R}^d$ a given compact set, $\{\phi_\Sigma(x - \mu), x \in C\}$ is uniformly equicontinuous as a family of functions of $(\mu, \Sigma)$ on $D$. This can be shown by adapting to the present context the arguments given in the last part of the proof of Theorem 2 of [31]. □

**Proof of Theorem 1.** According to Corollary 1 of [13], for any set of probability measures $\mathcal{Q}$ with $\inf\{d(F_0, Q) : Q \in \mathcal{Q}\} \geq 4\varepsilon$, any $\alpha, \beta > 0$ and all $n \geq 1$, there exists a test $\phi_n$ such that

$$F_0^n \phi_n \leq \sqrt{\frac{\beta}{\alpha}} N(\varepsilon, \mathcal{Q}, d) e^{-n\varepsilon^2}, \qquad \sup_{Q \in \mathcal{Q}} Q^n (1 - \phi_n) \leq \sqrt{\frac{\alpha}{\beta}} e^{-n\varepsilon^2}. \tag{24}$$

Let $A_\varepsilon = \{f : d(f_0, f) \leq 8\varepsilon\}$. Write

$$\Pi\left(A_\varepsilon^c | X_1, \ldots, X_n\right) = \Pi\left(A_\varepsilon^c \cap \mathcal{F}_n | X_1, \ldots, X_n\right) + \Pi\left(A_\varepsilon^c \cap \mathcal{F}_n^c | X_1, \ldots, X_n\right). \tag{25}$$

From Lemma 4 of [1], (4) implies that, for every $\eta > 0$

$$F_0^\infty \left\{ \int \prod_{i \leq n} (f(x_i)/f_0(x_i)) \, d\Pi(f) \leq \exp(-n\eta), \text{ i.o.} \right\} = 0, \tag{26}$$

so that, under (5), the second expression on the r.h.s. of (25) goes to 0 in $F_0^n$-probability (actually in $F_0^\infty$-almost surely). So it is sufficient to prove that the first expression in the r.h.s. of (25) goes to 0 in $F_0^n$-probability. Let $R_n(f) = \prod_{i \leq n} f(X_i)/f_0(X_i)$ and $I_n = \int R_n(f) \, d\Pi(f)$. The event $B_n$ that $I_n \geq e^{-cn\varepsilon^2}$ for $c$ in (6) satisfies $F_0^n(B_n) \to 1$, see (26). Therefore

$$F_0^n \left[ \Pi \left( A_\varepsilon^c \cap \mathcal{F}_n | X_1, \ldots, X_n \right) \right] = F_0^n \left[ \Pi \left( A_\varepsilon^c \cap \mathcal{F}_n | X_1, \ldots, X_n \right) \mathbb{1}_{B_n} \right] + o_p(1).$$

For arbitrary tests $\phi_{n,j}$, we have

$$F_0^n \left[ \Pi \left( A_\varepsilon^c \cap \mathcal{F}_{n,j} | X_1, \ldots, X_n \right) \mathbb{1}_{B_n} \right] \leq F_0^n \phi_{n,j} + F_0^n \left( (1 - \phi_{n,j}) \int_{A_\varepsilon^c \cap \mathcal{F}_{n,j}} R_n(f) \, d\Pi(f) \right) e^{cn\varepsilon^2}$$

$$\leq F_0^n \phi_{n,j} + \sup_{f \in \mathcal{F}_{n,j} : d(f_0, f) \geq 8\varepsilon} F^n (1 - \phi_{n,j}) \Pi(\mathcal{F}_{n,j}) e^{cn\varepsilon^2}$$

which can be bounded by a multiple of

$$\sqrt{\frac{\beta_j}{\alpha_j}} N(2\varepsilon, \mathcal{F}_{n,j}, d) e^{-4n\varepsilon^2} + \sqrt{\frac{\alpha_j}{\beta_j}} e^{-(4-c)n\varepsilon^2} \Pi(\mathcal{F}_{n,j})$$

for the choice of $\phi_{n,j}$ from Corollary 1 with $2\varepsilon$ in place of $\varepsilon$, $\mathcal{Q} = \{ F \in \mathcal{F}_{n,j} : d(f_0, f) \geq 8\varepsilon \}$ and any $\alpha_j, \beta_j > 0$. Put $\alpha_j = N(2\varepsilon, \mathcal{F}_{n,j}, d)$, $\beta_j = \Pi(\mathcal{F}_{n,j})$ to obtain

$$F_0^n \left[ \Pi \left( A_\varepsilon^c \cap \mathcal{F}_n | X_1, \ldots, X_n \right) \mathbb{1}_{B_n} \right] \leq \sum_j \sqrt{N(2\varepsilon, \mathcal{F}_{n,j}, d)} \sqrt{\Pi(\mathcal{F}_{n,j})} e^{-4n\varepsilon^2}$$

$$+ \sum_j \sqrt{N(2\varepsilon, \mathcal{F}_{n,j}, d)} \sqrt{\Pi(\mathcal{F}_{n,j})} e^{-(4-c)n\varepsilon^2}$$

so that (6) yields the result. □

**Lemma 2.** *Let $H, M \in \mathbb{N}$, $\underline{\sigma} > 0$ and, for $h = 1, \ldots, H$, $0 \leq \underline{a}_h < \bar{a}_h$, $1 \leq u_h$. Define*

$$\mathcal{G} = \left\{ f_P \text{ with } P = \sum_{h \geq 1} \pi_h \delta_{(\mu_h, \Sigma_h)} : \sum_{h > H} \pi_h \leq \varepsilon; \text{ for } h \leq H, \underline{a}_h < \|\mu_h\| \leq \bar{a}_h, \right.$$

$$\left. \underline{\sigma}^2 \leq \lambda_d(\Sigma_h), \lambda_1(\Sigma_h) \leq \underline{\sigma}^2 (1 + \varepsilon/\sqrt{d})^M, \lambda_1(\Sigma_h)/\lambda_d(\Sigma_h) \leq u_h \right\}.$$

*Then, for some positive constant $C_1$,*

$$N(\varepsilon, \mathcal{G}, \| \cdot \|_1) \lesssim \exp \left\{ dH \log M + H \log \frac{C_1}{\varepsilon} \right.$$

$$\left. + \sum_{h \leq H} \log \left[ \left( \frac{\bar{a}_h}{\underline{\sigma} \varepsilon/2} + 1 \right)^d - \left( \frac{\underline{a}_h}{\underline{\sigma} \varepsilon/2} - 1 \right)^d \right] + \frac{d(d-1)}{2} \log \frac{2du_h}{\varepsilon^2} \right\}. \tag{27}$$

**Proof.** The proof of (27) is based on a modification of the arguments of Proposition 2 in [27] to the location-scale mixture case. We recall here that a set $\hat{\mathcal{G}} \subset \mathcal{G}$ with the property that any element of $\mathcal{G}$ is within $\varepsilon$-distance from an element of $\hat{\mathcal{G}}$ is called an $\varepsilon$-*net* over $\mathcal{G}$. Since $N(\varepsilon, \mathcal{G}, d)$ is the minimal cardinality of an $\varepsilon$-net over $\mathcal{G}$, $N(\varepsilon, \mathcal{G}, d) \leq \#(\hat{\mathcal{G}})$.

Let $P_1 = \sum_{h \geq 1} \pi_h^{(1)} \delta_{(\mu_h^{(1)}, \Sigma_h^{(1)})}$ and $P_2 = \sum_{h \geq 1} \pi_h^{(2)} \delta_{(\mu_h^{(2)}, \Sigma_h^{(2)})}$. Then

$$
\begin{aligned}
&\|f_{P_1} - f_{P_2}\|_1 \\
&\leq \sum_{h \leq H} \pi_h^{(1)} \big\|\phi_{\Sigma_h^{(1)}}(\cdot - \mu_h^{(1)}) - \phi_{\Sigma_h^{(2)}}(\cdot - \mu_h^{(2)})\big\|_1 + \sum_{h \leq H} |\pi_h^{(1)} - \pi_h^{(2)}| + \sum_{h > H} \{\pi_h^{(1)} + \pi_h^{(2)}\}.
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\big\|\phi_{\Sigma_h^{(1)}}(\cdot - \mu_h^{(1)}) - \phi_{\Sigma_h^{(2)}}(\cdot - \mu_h^{(2)})\big\|_1 \\
&\leq \big\|\phi_{\Sigma_h^{(2)}}(\cdot - \mu_h^{(1)}) - \phi_{\Sigma_h^{(2)}}(\cdot - \mu_h^{(2)})\big\|_1 + \big\|\phi_{\Sigma_h^{(1)}} - \phi_{\Sigma_h^{(2)}}\big\|_1.
\end{aligned}
$$

The first term in the r.h.s. is smaller than $\sqrt{2/\pi}\|\mu_h^{(1)} - \mu_h^{(2)}\|/\sqrt{\lambda_d(\Sigma_h^{(2)})}$. As for the second term, use spectral decompositions $\Sigma_h^{(j)} = [O_h^{(j)} \Lambda_h^{(j)} (O_h^{(j)})^T]^{-1}$, $j = 1, 2$, where $O_h^{(1)}$, $O_h^{(2)}$ are orthogonal matrices and $\Lambda_h^{(1)}$, $\Lambda_h^{(2)}$ are diagonal matrices. Drop the index $h$ so to ease the notation and write $\Sigma_h^{(j)} = \Sigma_j$, $O_h^{(j)} = O_j$ and $\Lambda_h^{(j)} = \Lambda_j$. Moreover, let $\lambda_{j,1} \geq \cdots \geq \lambda_{j,d}$ be the diagonal elements of $\Lambda_j$. Finally, define $\widetilde{\Sigma} = (O_2 \Lambda_1 O_2^T)^{-1}$. By triangular inequality

$$
\|\phi_{\Sigma_1} - \phi_{\Sigma_2}\|_1 \leq \|\phi_{\widetilde{\Sigma}} - \phi_{\Sigma_2}\|_1 + \|\phi_{\widetilde{\Sigma}} - \phi_{\Sigma_1}\|_1. \tag{28}
$$

For each term in the r.h.s., use Csiszár's inequality, $\|f - g\|_1^2 \leq 2\int \log(f/g) f$, and the exact expression of the Kullback–Leibler divergence between zero mean multivariate Gaussians to get

$$
\int \log(\phi_{\Sigma_2}/\phi_{\Sigma_1}) \phi_{\Sigma_2} = \frac{1}{2}\big(\mathrm{tr}(\Sigma_1^{-1}\Sigma_2) - \log\det(\Sigma_1^{-1}\Sigma_2) - d\big).
$$

As for the first term in the r.h.s. of (28),

$$
\begin{aligned}
\|\phi_{\widetilde{\Sigma}} - \phi_{\Sigma_2}\|_1 &\leq \big\{\mathrm{tr}(\Sigma_2^{-1}\widetilde{\Sigma}) - \log\det(\Sigma_2^{-1}\widetilde{\Sigma}) - d\big\}^{1/2} \\
&= \left\{\sum_{i=1}^d (\lambda_{2,i}/\lambda_{1,i} - \log\lambda_{2,i}/\lambda_{1,i} - 1)\right\}^{1/2},
\end{aligned}
$$

where $\lambda_{2,i}/\lambda_{1,i}$ corresponds to $\lambda_{d-i+1}(\Sigma_h^{(1)})/\lambda_{d-i+1}(\Sigma_h^{(2)})$ in the original notation. As for the second term in the r.h.s. of (28),

$$
\begin{aligned}
\|\phi_{\widetilde{\Sigma}} - \phi_{\Sigma_1}\|_1 &\leq \big\{\mathrm{tr}(\Sigma_1^{-1}\widetilde{\Sigma}) - \log\det(\Sigma_1^{-1}\widetilde{\Sigma}) - d\big\}^{1/2} \\
&= \big\{\mathrm{tr}(Q\Lambda_1 Q^T \Lambda_1^{-1}) - d\big\}^{1/2} = \big\{\mathrm{tr}(Q\Lambda_1 Q^T \Lambda_1^{-1} - I)\big\}^{1/2},
\end{aligned}
$$

where $Q = O_2^T O_1$. Let $B = Q - I$ and $\|B\|_{\max} = \max |b_{i,j}|$. Then

$$\|B\|_{\max} \leq \|B\|_2 = \|O_2^T O_1 - I\|_2 \leq \|O_2^T\|_2 \|O_1 - O_2\|_2 = \|O_1 - O_2\|_2.$$

Let $b_{ij}$ be the element $(i, j)$ of $B$. Note that

$$\begin{aligned}
\mathrm{tr}(Q\Lambda_1 Q^T \Lambda_1^{-1} - I) &= \mathrm{tr}(B + \Lambda_1 B^T \Lambda_1^{-1} + B\Lambda_1 B^T \Lambda_1^{-1}) \\
&= \mathrm{tr}(B) + \mathrm{tr}(\Lambda_1 B^T \Lambda_1^{-1}) + \mathrm{tr}(B\Lambda_1 B^T \Lambda_1^{-1}) \\
&= \mathrm{tr}(B) + \mathrm{tr}(B^T) + \mathrm{tr}(B\Lambda_1 B^T \Lambda_1^{-1}) \\
&\leq 2\,\mathrm{tr}(B) + \frac{\lambda_{11}}{\lambda_{1d}} \mathrm{tr}(BB^T) \\
&= 2\,\mathrm{tr}(B) + 2\frac{\lambda_{11}}{\lambda_{1d}} \mathrm{tr}(I - Q) = 2\,\mathrm{tr}(B)\left[1 - \frac{\lambda_{11}}{\lambda_{1d}}\right] \\
&\leq 2d\|B\|_{\max}\left[\frac{\lambda_{1,1}}{\lambda_{1,d}} - 1\right] \leq 2d\|B\|_{\max}\frac{\lambda_{1,1}}{\lambda_{1,d}},
\end{aligned}$$

where in the second last inequality we use $\mathrm{tr}(B) \geq -d\|B\|_{\max}$ together with $\lambda_{1,1}/\lambda_{1,d} \geq 1$. Hence,

$$\|\phi_{\widetilde{\Sigma}} - \phi_{\Sigma_1}\|_1 \leq \left\{2d\|O_1 - O_2\|_2 \frac{\lambda_{1,1}}{\lambda_{1,d}}\right\}^{1/2},$$

where $\lambda_{1,1}/\lambda_{1,d}$ corresponds to $\lambda_1(\Sigma_h^{(1)})/\lambda_d(\Sigma_h^{(1)})$ in the original notation. In summary, back to the original notation,

$$\begin{aligned}
&\|f_{P_1} - f_{P_2}\|_1 \\
&\leq \sum_{h \leq H} \pi_h^{(1)} \left[\sqrt{\frac{2}{\pi}} \frac{\|\mu_h^{(1)} - \mu_h^{(2)}\|}{\sqrt{\lambda_d(\Sigma_h^{(2)})}} + \left\{\sum_{i=1}^d \left(\frac{\lambda_i(\Sigma_h^{(1)})}{\lambda_i(\Sigma_h^{(2)})} - \log\frac{\lambda_i(\Sigma_h^{(1)})}{\lambda_i(\Sigma_h^{(2)})} - 1\right)\right\}^{1/2} \right. \\
&\quad \left. + \left\{2d\|O_h^{(1)} - O_h^{(2)}\|_2 \frac{\lambda_1(\Sigma_h^{(1)})}{\lambda_d(\Sigma_h^{(1)})}\right\}^{1/2}\right] + \sum_{h \leq H} |\pi_h^{(1)} - \pi_h^{(2)}| + \sum_{h > H} \{\pi_h^{(1)} + \pi_h^{(2)}\}.
\end{aligned}$$

Now pick $f_P \in \mathcal{G}$ with $P = \sum_{h \geq 1} \pi_h \delta_{(\mu_h, \Sigma_h)}$, $\Sigma_h = (O_h \Lambda_h O_h^T)^{-1}$ and $\Lambda_h = \mathrm{diag}(\lambda_{h,1}, \ldots, \lambda_{h,d})$. Then find:

- $\hat{\mu}_h \in \hat{R}_h$, $h = 1, \ldots, H$, where $\hat{R}_h$ is a $\underline{\sigma}\varepsilon$-net of $R_h = \{\mu \in \mathbb{R}^d : \underline{a}_h < \|\mu\| \leq \bar{a}_h\}$ such that $\|\mu_h - \hat{\mu}_h\| < \underline{\sigma}\varepsilon$;
- $\hat{\pi}_1, \ldots, \hat{\pi}_H \in \hat{\Delta}$, where $\hat{\Delta}$ is a $\varepsilon$-net of the $H$-dimensional probability simplex $\Delta$ such that $\sum_{h \leq H} |\tilde{\pi}_h - \hat{\pi}_h| \leq \varepsilon$, and $\tilde{\pi}_h = \pi_h / \sum_{l \leq H} \pi_l$, $h \leq H$;
- $\hat{O}_h \in \hat{\mathcal{O}}_h$, $h = 1, \ldots, H$, where $\hat{\mathcal{O}}_h$ is a $\delta_h$-net of the set $\mathcal{O}$ of $d \times d$ orthogonal matrices with respect to the spectral norm $\|\cdot\|_2$, with $\delta_h = \varepsilon^2/(2du_h)$ such that $\|O_h - \hat{O}_h\|_2 \leq \delta_h$;

- $(m_{h,1}, \ldots, m_{h,d}) \in \{1, \ldots, M\}^d$, $h = 1, \ldots, H$, such that $\hat{\lambda}_{h,i} = \{\underline{\sigma}^2(1 + \varepsilon/\sqrt{d})^{m_{h,i}-1}\}^{-1}$ satisfies $1 \le \hat{\lambda}_{h,i}/\lambda_{h,i} < (1 + \varepsilon/\sqrt{d})$.

Take $\hat{P} = \sum_{h \le H} \hat{\pi}_h \delta_{(\hat{\mu}_h, \hat{\Sigma}_h)}$ and $\hat{\Sigma}_h = (\hat{O}_h \hat{\Lambda}_h \hat{O}_h^T)^{-1}$, where $\hat{\Lambda}_h = \mathrm{diag}(\hat{\lambda}_{h,1}, \ldots, \hat{\lambda}_{h,d})$. Then

$$\|f_P - f_{\hat{P}}\|_1 \le \max_{h \le H}\left[\sqrt{\frac{2}{\pi}}\frac{\|\mu_h - \hat{\mu}_h\|}{\sqrt{\lambda_d(\hat{\Sigma}_h)}} + \left\{\sum_{i=1}^d \left(\frac{\lambda_i(\Sigma_h)}{\lambda_i(\hat{\Sigma}_h)} - \log\frac{\lambda_i(\Sigma_h)}{\lambda_i(\hat{\Sigma}_h)} - 1\right)\right\}^{1/2}\right]$$

$$+ \sum_{h \le H} \pi_h \left\{2d\|O_h - \hat{O}_h\|_2 \frac{\lambda_1(\Sigma_h)}{\lambda_d(\hat{\Sigma}_h)}\right\}^{1/2} + \sum_{h \le H} |\pi_h - \hat{\pi}_h| + \sum_{h > H} \pi_h$$

$$\le \max_{h \le H}\left[\sqrt{\frac{2}{\pi}}\frac{\|\mu_h - \hat{\mu}_h\|}{\underline{\sigma}} + \left\{\sum_{i=1}^d \left(\frac{\hat{\lambda}_{h,d-i+1}}{\lambda_{h,d-i+1}} - 1\right)^2\right\}^{1/2}\right]$$

$$+ \sum_{h \le H} \pi_h \left\{2d\|O_h - \hat{O}_h\|_2 \frac{\lambda_{h,1}}{\lambda_{h,d}}\right\}^{1/2} + \sum_{h \le H} |\pi_h - \hat{\pi}_h| + \sum_{h > H} \pi_h,$$

where in the second inequality $x - \log x - 1 \le (x - 1)^2$ for $x \ge 1$ together with $\hat{\lambda}_{h,d-i+1}/\lambda_{h,d-i+1} \ge 1$ has been used to bound the second term. Moreover,

$$\sum_{h \le H} |\pi_h - \hat{\pi}_h| \le \sum_{h \le H}\left|\pi_h - \left(1 - \sum_{h > H}\pi_h\right)\hat{\pi}_h\right| + \sum_{h \le H}\left|\left(1 - \sum_{h > H}\pi_h\right)\hat{\pi}_h - \hat{\pi}_h\right|$$

$$= \left(1 - \sum_{h > H}\pi_h\right)\sum_{h \le H}|\tilde{\pi}_h - \hat{\pi}_h| + \left(\sum_{h > H}\pi_h\right)\sum_{h \le H}\hat{\pi}_h$$

$$\le \sum_{h \le H}|\tilde{\pi}_h - \hat{\pi}_h| + \sum_{h > H}\pi_h \le 2\varepsilon.$$

Hence,

$$\|f_P - f_{\hat{P}}\|_1 \le \sqrt{\frac{2}{\pi}}\varepsilon + \left\{\sum_{i=1}^d \left(1 + \frac{\varepsilon}{\sqrt{d}} - 1\right)^2\right\}^{1/2} + \sum_{h \le H}\pi_h\left\{2d\frac{\varepsilon^2}{2du_h}\frac{\lambda_{h,1}}{\lambda_{h,d}}\right\}^{1/2} + 2\varepsilon + \varepsilon$$

$$\le \varepsilon + \{\varepsilon^2\}^{1/2} + \{\varepsilon^2\}^{1/2} + 3\varepsilon = 6\varepsilon.$$

Thus a $6\varepsilon$-net of $\mathcal{G}$, in the $L_1$ distance, can be constructed with $f_{\hat{P}}$ as above. Recalling that $\#(\hat{\Delta}) \lesssim \varepsilon^{-H}$, $\#(\hat{O}_h) \lesssim \delta_h^{-d(d-1)/2}$ and $\#(\hat{R}_h) \le (\bar{a}_h/(\underline{\sigma}\varepsilon/2) + 1)^d - (\underline{a}_h/(\underline{\sigma}\varepsilon/2) - 1)^d$, the total number is

$$\lesssim (M)^{dH}\varepsilon^{-H}\prod_{h \le H}\left[\left(\frac{\bar{a}_h}{\underline{\sigma}\varepsilon/2} + 1\right)^d - \left(\frac{\underline{a}_h}{\underline{\sigma}\varepsilon/2} - 1\right)^d\right]C_1\left(\frac{2du_h}{\varepsilon^2}\right)^{d(d-1)/2}$$

for some positive constant $C_1$. Finally, the constant factor by 6 can be absorbed in the bound, hence (27) follows. □

**Remark 2.** The usual low-entropy, high mass sieve approach of Theorem 2 by [11] is not suitable for multivariate DP location-scale mixture. This theorem, indeed, requires a sieve $\mathcal{F}_n$ such that: (i) $\Pi(\mathcal{F}_n^c)$ decreases exponentially fast in $n$, and (ii) the $L_1$ metric entropy of $\mathcal{F}_n$ grows linearly with $n$. As for (i), let $\mathcal{F}_n$ to be defined as $\mathcal{G}$ of Lemma 2 with $M = \underline{\sigma}^{-2c_2} = \bar{a}_h = u_h = n$, $H_n = \lfloor Cn\varepsilon^2/\log n \rfloor$, $\underline{a}_h = 0$ and $l_h = 1$. Clearly $\mathcal{F}_n \uparrow \mathcal{F}$ as $n \to \infty$. However, in order to have $\mathcal{F}_n$ satisfying condition (i), stronger tail conditions on $P^*$ are necessary than those in (7) and (10), namely exponential tail behaviors. While one can be still satisfied with a Gaussian specification for $\mu$, exponential tail for the condition number turns out to be too restrictive ruling out all the models discussed in Section 3 as well as any other reasonable prior specification for non diagonal covariance matrices. However, as an inspection of the proof of Lemma 2 reveals, with diagonal covariance matrices the metric entropy is considerably smaller and one can prove consistency by applying Theorem 2 of [11] under essentially the same tail requirements of DP location mixture of [27]. We omit the details here.

**Proof of Corollary 1.** That conditions (8) and (9) are satisfied when $\mathcal{L} = \mathrm{IW}(\Sigma_0, \nu)$ has been proved in Lemma 1 of [27]. Therefore we focus here on condition (10) on the condition number. Let for the moment $\Sigma_0 = I$. We directly work with the joint distribution of the ordered eigenvalues of a Wishart-distributed matrix:

$$f_{\lambda_1,\ldots,\lambda_d}(x_1,\ldots,x_d) = K \exp\left\{-\frac{1}{2}\sum_{i=1}^{d} x_i\right\} \prod_{j=1}^{d} (x_j)^{(\nu-d-1)/2} \prod_{i<j}(x_i - x_j),$$

over the set $\{(x_1,\ldots,x_d) \in (0,\infty)^d : x_1 \geq \cdots \geq x_d\}$, where $K$ is a known normalizing constant. See Corollary 3.2.19, page 107, of [24]. To show condition (10), let $z = \lambda_1/\lambda_d$ and apply the Jacobi's transformation formula for $g(\lambda_1,\ldots,\lambda_d) = (z,\lambda_2,\ldots,\lambda_d)$. For $J_{g^{-1}}$ denoting the Jacobian of $g^{-1}$, we have

$$f_{z,\lambda_2,\ldots,\lambda_d}(z,\ldots,y) = |J_{g^{-1}}| f_{\lambda_1,\ldots,\lambda_d}(x_d z, x_2,\ldots,x_d),$$

and the marginal of $z$ can be obtained integrating out $x_2,\ldots,x_d$. Using $|J_{g^{-1}}| = x_d$, $x_d \geq 0$ and $x_d \leq x_i \leq zx_d$ for $i = 2,\ldots,d-1$, after some algebra one gets

$$f_z(z) \leq K'' z^{-(\nu-d+3)/2}$$

for some constant $K''$, so that $P^*(z > x) \lesssim x^{-(\nu-d+1)/2}$ as $x \to \infty$. Cfr. Theorem 3.2 of [8]. If $\Sigma_0 \neq I$, the conclusion holds for a different set of constants, see Theorem 4 of [23] for the complex Wishart case. Condition (10) is then satisfied under the hypothesis on $\nu$. □

**Proof of Corollary 2.** To prove condition (8), consider

$$P^*\{\lambda_1(\Sigma^{-1}) \geq x\} \leq P^*\{\mathrm{tr}(\Sigma^{-1}) \geq x\} \leq P^*\{\mathrm{tr}(\Omega^{-1}) \geq x\},$$

where the last inequality is verified since $\text{tr}(\Sigma^{-1}) \leq \text{tr}(\Omega^{-1})$ by an application of the Woodbury's identity. Since the trace of $\Omega^{-1}$ is the sum of i.i.d. gamma random variable, and the gamma has exponential tail, (8) is verified. As for (9), consider

$$P^*\{\lambda_d(\Sigma^{-1}) < 1/x\} \leq x^{-1} E_{P^*}\{\lambda_1(\Sigma)\} \leq x^{-1}[E_{P^*}\{\lambda_1(\Omega)\} + E_{P^*}\{\lambda_1(\Gamma\Gamma^T)\}]$$

by Markov's and Weyl's inequalities. The expectation of $\lambda_1(\Omega)$ is finite and since $\lambda_1(\Gamma\Gamma^T) \leq \text{tr}(\Gamma\Gamma^T) \sim \chi^2_{rd}$, also the expectation of $\lambda_1(\Gamma\Gamma^T)$ is finite. Finally, to prove condition (10), first use Markov's inequality. For $k > d(d-1)$, we have

$$P^*\{\lambda_1(\Sigma^{-1})/\lambda_d(\Sigma^{-1}) \geq x\} = P^*\{(\lambda_1(\Sigma)/\lambda_d(\Sigma))^k \geq x^k\}$$
$$\leq x^{-k} E\{(\lambda_1(\Sigma)/\lambda_d(\Sigma))^k\}.$$

Using again Weyl's inequalities and noting that $\lambda_d(\Gamma\Gamma^T) = 0$ when $r < d$, we have

$$\frac{\lambda_1(\Sigma)}{\lambda_d(\Sigma)} \leq \frac{\lambda_1(\Omega) + \lambda_1(\Gamma\Gamma^T)}{\lambda_d(\Omega) + \lambda_d(\Gamma\Gamma^T)} = \frac{\lambda_1(\Omega) + \lambda_1(\Gamma\Gamma^T)}{\lambda_d(\Omega)} \leq \frac{\lambda_1(\Omega^{-1})}{\lambda_d(\Omega^{-1})} + \frac{\text{tr}(\Gamma\Gamma^T)}{\lambda_d(\Omega)}.$$

By convexity of $g(x) = x^k$, for any $x, y \in \mathbb{R}^+$ we have $(x + y)^k \leq 2^{k-1}(x^k + y^k)$. Hence, $\lambda_1(\Sigma)/\lambda_d(\Sigma)$ has finite $k$th moment if both summand have finite $k$th moment. First, consider the distribution of $\lambda_1(\Omega^{-1})/\lambda_d(\Omega^{-1})$. Since $\Omega^{-1}$ is a diagonal matrix with i.i.d. gamma distributed diagonal elements, the joint distribution of the ordered eigenvalues of $\Omega^{-1}$ is the joint distribution of the ordered statistics of an i.i.d. sample from $\text{Ga}(a, b)$, that is,

$$f_{\lambda_1(\Omega^{-1})\cdots\lambda_d(\Omega^{-1})}(x_1, \ldots, x_d) = d! f_{\sigma_1^{-2}}(x_1) \times \cdots \times f_{\sigma_d^{-2}}(x_d) = K \prod_{j=1}^{d} x_j^{a-1} \exp\left\{-b\sum_{j=1}^{d} x_j\right\},$$

over the set $\{(x_1, \ldots, x_d) \in (0, \infty)^d : x_1 \geq \cdots \geq x_d\}$, where $K$ is a known normalizing constant. Now we use the Jacobi's transformation formula as in the proof of Corollary 1 to get, for $z = \lambda_1(\Omega^{-1})/\lambda_d(\Omega^{-1})$,

$$f_z(z) = K \int_{x_3}^{\infty} \cdots \int_{x_{d-2}}^{\infty} \int_{x_d}^{\infty} \prod_{j=2}^{d-1} x_j^{a-1} \exp\left\{-b\sum_{j=1}^{d} x_j\right\} dx_2 \cdots dx_{d-1}$$

$$\times \int_0^{\infty} x_d(zx_d^2)^{a-1} \exp\{-bx_d(z+1)\} dx_d \leq \cdots \leq K' \frac{z^{a-1}}{(z+1)^{2a}}.$$

Hence, $\lambda_1(\Omega^{-1})/\lambda_d(\Omega^{-1})$ has finite $k$th moment as long as $a > k$. Now, since

$$\frac{\text{tr}(\Gamma\Gamma^T)}{\lambda_d(\Omega)} = \text{tr}(\Gamma\Gamma^T)\lambda_1(\Omega^{-1})$$

and the $k$th moment of the product of independent random variables is the product of the $k$th moments, it remains to show that both $\text{tr}(\Gamma\Gamma^T)$ and $\lambda_1(\Omega^{-1})$ have finite $k$th moment. To this aim, it is sufficient to note that $\text{tr}(\Gamma\Gamma^T) \sim \chi^2_{dr}$ and $\lambda_1(\Omega^{-1})$ has the distribution of the

first order statistics of a sample of $d$ independent gamma random variables. The proof is then complete. □

**Proof of Corollary 3.** Since each eigenvalues has independent inverse-gamma distribution, conditions (8)–(10) are satisfied following part of the proof of Corollary 2. □

**Proof of Proposition 1.** Let $M_n = \underline{\sigma}_n^{-2c_2} = n$ with $\varepsilon_n = n^{-\gamma}(\log n)^t$ for $t_0 = t + s$ and $H_n = \lfloor Cn\varepsilon_n^2/\log n \rfloor$ for a positive constant to be determined later. Also define $\mathcal{F}_n$ and $\mathcal{F}_{n,\mathbf{j},\mathbf{l}}$ as in the proof of Theorem 2 with $\varepsilon = \varepsilon_n$. We prove next that

$$\Pi(\mathcal{F}_n^c) \lesssim e^{-4n\tilde{\varepsilon}_n^2}, \tag{29}$$

$$\sum_{\mathbf{j},\mathbf{l}} \sqrt{N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\mathbf{l}}, d)}\sqrt{\Pi(\mathcal{F}_{n,\mathbf{j},\mathbf{l}})}e^{-n\varepsilon_n^2} \to 0. \tag{30}$$

The thesis then follows by an application of Theorem 5 by [13], cfr. version in Theorem 3 of [18].

As for (29), refer to equation (11). Note that $n\varepsilon_n^2 = n^{1-2\gamma}(\log n)^{2t_0}$ and $H_n = C \times n^{1-2\gamma}(\log n)^{2t_0-1}$. We have

$$\left\{ \frac{e\alpha}{H_n}\log\frac{1}{\varepsilon_n} \right\}^{H_n} \lesssim \exp\{-H_n \log H_n\}$$

$$= \exp\{-Cn^{1-2\gamma}(\log n)^{2t_0-1}\log(Cn^{1-2\gamma}(\log n)^{2t_0-1})\}$$

$$\lesssim \exp\{-Cn^{1-2\gamma}(\log n)^{2t_0-1}(1-2\gamma)\log n\}$$

$$= \exp\{-(1-2\gamma)Cn\varepsilon_n^2\},$$

$$H_n e^{-c_1\underline{\sigma}_n^{-2c_2}} = Cn^{1-2\gamma}(\log n)^{2t_0-1}e^{-nc_1} = o(\exp\{-n\varepsilon_n^2\}),$$

$$H_n\underline{\sigma}_n^{-2c_3}(1+\varepsilon_n/\sqrt{d})^{-c_3 M_n} \leq Cn^{c_3/c_2+1-2\gamma}(\log n)^{2t_0-1}\exp\{-c_3 n\varepsilon_n/(2\sqrt{d})\}$$

$$= o(\exp\{-n\varepsilon_n^2\}).$$

Therefore, $\Pi(\mathcal{F}_n^c) \lesssim \exp\{-(1-2\gamma)Cn\varepsilon_n^2\} \lesssim e^{-4n\tilde{\varepsilon}_n^2}$ for all large $n$ since $t < t_0$. Note that the rate at which the prior probability of the complement of the sieve vanishes is determined by the truncation level of the stick-breaking weights.

As for (30), refer to equation (12). Based on the inequality $d(f,g)^2 \leq \|f-g\|_1$,

$$N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\mathbf{l}}, d) \lesssim \exp\left\{dH_n \log n + H_n \log\frac{C_1}{\varepsilon_n^2} + \sum_{h \leq H_n}\log\frac{n^{c_4 d}j_h^{d-1}}{\varepsilon_n^{2d}} + \frac{d(d-1)}{2}\log\frac{2dn^{2^{l_h}}}{\varepsilon_n^4}\right\}.$$

Since $H_n \log(1/\varepsilon_n) = \gamma Cn\varepsilon_n^2 + o(n\varepsilon_n^2)$, we also have

$$N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\mathbf{l}}, d) \lesssim \exp\{[d+c_4 d+2\gamma+2\gamma d+2\gamma d(d-1)]Cn\varepsilon_n^2\}\prod_{h \leq H_n}j_h^{d-1}n^{(d(d-1)/2)2^{l_h}}$$

$$< \exp\{d^* Cn\varepsilon_n^2\}\prod_{h \leq H_n}j_h^{d-1}n^{(d(d-1)/2)2^{l_h}},$$

where $d^* = d + c_4 d + 1 + d + d(d-1)$ and $\gamma < 1/2$ has been used in the last inequality. By using (13), $\sqrt{N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\mathbf{l}}, d)}\sqrt{\Pi(\mathcal{F}_{n,\mathbf{j},\mathbf{l}})}$ can be bounded by a multiple of

$$\exp\{(d^*/2)Cn\varepsilon_n^2\} \prod_{h \leq H_n} j_h^{(d-1)/2}\big[\sqrt{n}(j_h - 1)\big]^{-\mathbb{1}_{(j_h \geq 2)}(r+1)} n^{(d(d-1)/4)2^{l_h} - \mathbb{1}_{(l_h \geq 1)}(\kappa/2)2^{l_h-1}}.$$

Summing with respect to $\mathbf{j}$ and $\mathbf{l}$, by arguments similar to those used in the proof of Theorem 2, we obtain, for $n$ large enough and when $r > (d-1)/2$ and $\kappa > d(d-1)$, the upper bound

$$\sum_{\mathbf{j},\mathbf{l}} \sqrt{N(\varepsilon_n, \mathcal{F}_{n,\mathbf{j},\mathbf{l}}, d)}\sqrt{\Pi(\mathcal{F}_{n,\mathbf{j},\mathbf{l}})}$$

$$\leq \exp\{(d^*/2)Cn\varepsilon_n^2\}\big(1 + n^{-(r+1)/2}K\big)^{H_n}\big[2n^{d(d-1)/4}\big]^{H_n}$$

$$\lesssim \exp\{(d^*/2 + d(d-1)/4)Cn\varepsilon_n^2\},$$

where $K := \sum_{j \geq 2} j^{(d-1)/2}(j-1)^{-(r+1)} < \infty$. Hence, (30) is satisfied for $C$ sufficiently small to satisfy $C \leq 4/[2d^* + d(d-1)]$. The proof is complete. $\square$

**Proof of Proposition 2.** The proof follows the arguments presented in Theorem 4 of [27] adapted to the location-scale case. Denote by $F_0 g = \int g(x) f_0(x)\,\mathrm{d}x$ the expectation of $g(X)$ under $X \sim f_0$. Consider $n$ large enough such that $\tilde{\varepsilon}_n < s_0^\beta$ and fix $\sigma_n^\beta = \tilde{\varepsilon}_n\{\log(1/\tilde{\varepsilon}_n)\}^{-1}$. As in Proposition 1 of [27], define $E_{\sigma_n} = \{\|x\| \in \mathbb{R}^d : f_0(x) \geq \sigma_n^{(4\beta+2\varepsilon+8)/\delta}\}$ for some $\delta \in (0, 1)$ and $a_{\sigma_n} = a_0\{\log(1/\sigma_n)\}^{1/\tau}$. Then, $F_0(E_{\sigma_n}^c) \leq B_0\sigma_n^{4\beta+2\varepsilon+8}$ for some constant $B_0$, $E_{\sigma_n} \subset \{\|x\| \in \mathbb{R}^d : \|x\| \leq a_{\sigma_n}\}$ and, there exists a density $\tilde{h}_{\sigma_n}$ with support $E_{\sigma_n}$ such that $d(f_0, \phi_{\sigma_n} \star \tilde{h}_{\sigma_n}) \lesssim \sigma_n^\beta$. Find $b_1 > \max\{1, 1/(2\beta)\}$ such that $\tilde{\varepsilon}_n^{b_1}\{\log(1/\varepsilon_n)\}^{5/4} \leq \tilde{\varepsilon}_n$. Apply Theorem B1 of [27] for $\sigma = \sigma_n$, $a = a_{\sigma_n}$, $F = \tilde{h}_{\sigma_n}$ and $\varepsilon = \tilde{\varepsilon}_n^{2b_1}$ to get $F_{\sigma_n} = \sum_{j=1}^N p_j\delta_{\mu_j}$ with

$$N \leq D\big[\big(a_{\sigma_n}\sigma_n^{-1} \vee 1\big)\log\big(1/\tilde{\varepsilon}_n^{2b_1}\big)\big]^d \lesssim \sigma_n^{-d}\{\log(1/\tilde{\varepsilon}_n)\}^{d+d/\tau}$$

many support points inside $\{x \in \mathbb{R} : \|x\| \leq a_{\sigma_n}\}$ such that

$$\|\phi_{\sigma_n} \star \tilde{h}_{\sigma_n} - \phi_{\sigma_n} \star F_{\sigma_n}\|_1 \lesssim \tilde{\varepsilon}_n^{2b_1}\big\{\log\big(1/\tilde{\varepsilon}_n^{2b_1}\big)\big\}^{1/2}.$$

Move each of the support point of $F_{\sigma_n}$ to the nearest point on the grid $\{(n_1, \ldots, n_d)\sigma_n\tilde{\varepsilon}_n^{2b_1} : n_i \in \mathbb{Z}, |n_i| < \lceil a_{\sigma_n}/(\sigma_n\tilde{\varepsilon}_n^{2b_1})\rceil, i = 1, \ldots, d\}$, so that $\min_{j \neq l}\|\mu_j - \mu_l\| > \sigma_n\tilde{\varepsilon}_n^{2b_1}$. These moves cost at most a constant times $\tilde{\varepsilon}_n^{2b_1}$ to the $L_1$ distance. Modify further $F_{\sigma_n}$ so that $(\mu_1, \ldots, \mu_N)$ form a $\sigma_n a_{\sigma_n}^{\tau/2}$-net of $\{x \in \mathbb{R}^d : \|x\| \leq a_{\sigma_n}\}$. This can be achieved by adding at most $[a_{\sigma_n}/(\sigma_n a_{\sigma_n}^{\tau/2}) + 1]^d \lesssim \sigma_n^{-d}a_{\sigma_n}^{d-d\tau/2} \lesssim \sigma_n^{-d}\{\log(1/\sigma_n)\}^{d/\tau-d/2}$ support points so that we still have $N \lesssim \sigma_n^{-d}\{\log(1/\tilde{\varepsilon}_n)\}^{d+d/\tau}$. Finally, modify the probability masses $p_1, \ldots, p_N$ to a version for which $p_j \geq \tilde{\varepsilon}_n^{4db_1}$ for $j = 1, \ldots, N$. By a suitable extension of Lemma 3 of [18] to $d$ dimensions, the $L_1$ distance changes by the $L_1$ distance between the original and the new probability weights,

and the latter can be shown to be smaller than $2(N-1)\tilde{\varepsilon}_n^{4db_1}$. In conclusion, we have

$$
\begin{aligned}
\|\phi_{\sigma_n} \star \tilde{h}_{\sigma_n} - \phi_{\sigma_n} \star F_{\sigma_n}\|_1 &\lesssim \tilde{\varepsilon}_n^{2b_1}\left\{\log\left(1/\tilde{\varepsilon}_n^{2b_1}\right)\right\}^{1/2} + \tilde{\varepsilon}_n^{2b_1} + N\tilde{\varepsilon}_n^{4db_1} \\
&\lesssim \tilde{\varepsilon}_n^{2b_1}\left\{\log(1/\tilde{\varepsilon}_n)\right\}^{1/2} + \sigma_n^{-d}\left\{\log(1/\tilde{\varepsilon}_n)\right\}^{d+d/\tau}\tilde{\varepsilon}_n^{4db_1} \\
&= \tilde{\varepsilon}_n^{2b_1}\left\{\log(1/\tilde{\varepsilon}_n)\right\}^{1/2} + \tilde{\varepsilon}_n^{d(4b_1-1/\beta)}\left\{\log(1/\tilde{\varepsilon}_n)\right\}^{1+1/\tau+1/\beta} \\
&\lesssim \tilde{\varepsilon}_n^{2b_1}\left\{\log(1/\tilde{\varepsilon}_n)\right\}^{1/2}
\end{aligned}
\tag{31}
$$

since $d(4b_1 - 1/\beta) > 2b_1$ for $b_1 > 1/(4d-2)\beta$ and the latter follows by the assumption made on $b_1$.

Place disjoint balls $U_j$ with centers at $\mu_1, \ldots, \mu_N$ with diameter $\sigma_n\tilde{\varepsilon}_n^{2b_1}$ each. Because of assumption (20) on $f^*$, for any $\|x\| \leq a_{\sigma_n}$,

$$
f^*(x) \geq c^* e^{-b^* a_{\sigma_n}^{\tau^*}} \geq c^* \exp\left\{-b^* a_0^{\tau^*}\log(1/\sigma_n)^{\tau^*/\tau}\right\} \geq c^* \sigma_n^{b^* a_0^{\tau^*}\log(1/\sigma_n)^{\tau^*/\tau-1}} \gtrsim \sigma_n^{b^* a_0^{\tau^*}},
$$

where the second inequality follows from the assumption that $\tau^* \leq \tau$. This implies that for $n$ sufficiently large and some constant $a_1$, $a_1\sigma_n^{b^* a_0^{\tau^*}}(\sigma_n\tilde{\varepsilon}_n^{2b_1})^d \leq P^*(U_j) \leq 1$. Further extend this to a partition $U_1, \ldots, U_K$ of $\mathbb{R}^d$ such that $a_1\sigma_n^{b^* a_0^{\tau^*}}(\sigma_n\tilde{\varepsilon}_n^{2b_1})^d \leq P^*(U_j) \leq 1$ still holds. We can still have $K \lesssim \sigma_n^{-d}\{\log(1/\tilde{\varepsilon}_n)\}^{d+d/\tau} = \tilde{\varepsilon}_n^{-d/\beta}\{\log(1/\tilde{\varepsilon}_n)\}^{sd}$ for $s = 1 + 1/\beta + 1/\tau$. The next step is to construct a partition of $\mathbb{R}^d \times \mathcal{S}$. To this aim, let

$$
S_{\sigma_n} = \left\{\Sigma \in \mathcal{S} : \sigma_n^{-2} < \lambda_i\left(\Sigma^{-1}\right) < \sigma_n^{-2}\left(1 + \sigma_n^{2\beta}\right), i = 1, \ldots, d\right\}
$$

which satisfies $P^*(S_{\sigma_n}) \gtrsim e^{-C_3\sigma_n^{-\kappa^*}}$ by hypothesis (19). Define $V_j = U_j \times S_{\sigma_n}$ for $j = 1, \ldots, K$, and extend $\{V_1, \ldots, V_K\}$ to a partition $\{V_1, \ldots, V_M\}$ of $\mathbb{R}^d \times \mathcal{S}$ such that

$$
a_2\sigma_n^{b^* a_0^{\tau^*}+d}\sigma_n\tilde{\varepsilon}_n^{2db_1}e^{-C_3\sigma_n^{-\kappa^*}} \leq \alpha P^*(V_j) \leq 1
$$

for all $j = 1, \ldots, M$, for some positive constant $a_2$ and $n$ sufficiently large. We can still have $M \lesssim \tilde{\varepsilon}_n^{-d/\beta}\{\log(1/\tilde{\varepsilon}_n)\}^{sd}$.

Set $p_j = 0$ for $j > N$ and consider the set $\mathcal{P}_{\sigma_n} \subset \mathcal{P}$ of probability measures $P$ on $\mathbb{R}^d \times \mathcal{S}$ with

$$
\sum_{j=1}^{M}\left|P(V_j) - p_j\right| \leq 2\tilde{\varepsilon}_n^{2db_1}e^{-C_3\sigma_n^{-\kappa^*}}.
$$

Note that

$$
M\tilde{\varepsilon}_n^{2db_1}e^{-C_3\sigma_n^{-\kappa^*}} \lesssim \tilde{\varepsilon}_n^{-d/\beta}\left\{\log(1/\tilde{\varepsilon}_n)\right\}^{sd}\tilde{\varepsilon}_n^{2db_1}e^{-C_3\sigma_n^{-\kappa^*}} \leq 1,
$$

$$
\begin{aligned}
\min_{1 \leq j \leq M}\left[\alpha P^*(V_j)\right]^{1/3} &\geq \left\{a_2\sigma_n^{b^* a_0^{\tau^*}+d}\tilde{\varepsilon}_n^{2db_1}e^{-C_3\sigma_n^{-\kappa^*}}\right\}^{1/3} \\
&= a_2^{1/3}\tilde{\varepsilon}_n^{2db_1}e^{-C_3\sigma_n^{-\kappa^*}}\left\{\sigma_n^{b^* a_0^{\tau^*}+d}\tilde{\varepsilon}_n^{-4db_1}e^{2C_3\sigma_n^{-\kappa^*}}\right\}^{1/3} \geq a_2^{1/3}\tilde{\varepsilon}_n^{2db_1}e^{-C_3\sigma_n^{-\kappa^*}}
\end{aligned}
$$

for $n$ sufficiently large. Hence, by Lemma 10 of [13],

$$
\begin{aligned}
\mathcal{D}_{\alpha, P^*}(\mathcal{P}_{\sigma_n}) &\geq C' \exp\big\{-cM \log\big[1/\tilde{\varepsilon}_n^{2db_1} e^{-C_3 \sigma_n^{-\kappa^*}}\big]\big\} \geq C'' \exp\big\{-c'M \sigma_n^{-\kappa^*}\big\} \\
&\geq C'' \exp\big\{-c'' \tilde{\varepsilon}_n^{-d/\beta} \big\{\log(1/\tilde{\varepsilon}_n)\big\}^{sd} \sigma_n^{-\kappa^*}\big\} \\
&= C'' \exp\big\{-c''' \tilde{\varepsilon}_n^{-(d+\kappa^*)/\beta} \big\{\log(1/\tilde{\varepsilon}_n)\big\}^{d(s+\kappa^*/\beta)}\big\} \geq \exp\big\{-Cn\tilde{\varepsilon}_n^2\big\}
\end{aligned}
\tag{32}
$$

for some constants $C, C', C'', c, c', c'', c'''$ and for $t \geq d(s + \kappa^*/\beta)/(2 + (d + \kappa^*)/\beta)$ since

$$
\begin{aligned}
&\tilde{\varepsilon}_n^{-(d+\kappa^*)/\beta} \big\{\log(1/\tilde{\varepsilon}_n)\big\}^{d(s+\kappa^*/\beta)} \\
&= n^{(d+\kappa^*)/(2\beta+d+\kappa^*)} (\log n)^{-(d+\kappa^*)t/\beta} \big\{\log\big[n^{\beta/(2\beta+d+\kappa^*)} (\log n)^{-t}\big]\big\}^{d(s+\kappa^*/\beta)} \\
&\lesssim n^{(d+\kappa^*)/(2\beta+d+\kappa^*)} (\log n)^{d(s+\kappa^*/\beta)-(d+\kappa^*)t/\beta} \leq n\tilde{\varepsilon}_n^2
\end{aligned}
$$

when $d(s + \kappa^*/\beta) - (d + \kappa^*)t/\beta \leq 2t$, cfr. (22).

We show next that, for any $P \in \mathcal{P}_{\sigma_n}$, $d(f_0, f_P) \lesssim \sigma_n^\beta$. By triangle inequality,

$$
d(f_0, f_P) \leq d(f_0, \phi_{\sigma_n} \star \tilde{h}_{\sigma_n}) + d(\phi_{\sigma_n} \star \tilde{h}_{\sigma_n}, \phi_{\sigma_n} \star F_{\sigma_n}) + d(f_P, \phi_{\sigma_n} \star F_{\sigma_n})
$$

and $d(f_0, \phi_{\sigma_n} \star \tilde{h}_{\sigma_n}) \lesssim \sigma_n^\beta$ by Proposition 1 of [27], while, by (31),

$$
d(\phi_{\sigma_n} \star \tilde{h}_{\sigma_n}, \phi_{\sigma_n} \star F_{\sigma_n}) \lesssim \tilde{\varepsilon}_n^{b_1} \big\{\log(1/\tilde{\varepsilon}_n)\big\}^{1/4} \leq \tilde{\varepsilon}_n \big\{\log(1/\tilde{\varepsilon}_n)\big\}^{-1} = \sigma_n^\beta
$$

since, by assumption, $\tilde{\varepsilon}_n^{b_1} \{\log(1/\varepsilon_n)\}^{5/4} \leq \tilde{\varepsilon}_n$. In order to show that also $d(f_P, \phi_{\sigma_n} \star F_{\sigma_n}) \lesssim \sigma_n^\beta$, we use arguments presented in Lemma 5 of [13]. Let $V_0 = \bigcup_{j>N} V_j$. After some algebra,

$$
\begin{aligned}
&f_P(x) - \phi_{\sigma_n} \star F_{\sigma_n}(x) \\
&= \int_{V_0} \phi_\Sigma(x - \mu) \, dP(\mu, \Sigma) + \sum_{j=1}^N \int_{V_j} \big(\phi_\Sigma(x - \mu) - \phi_\Sigma(x - \mu_j)\big) \, dP(\mu, \Sigma) \\
&\quad + \sum_{j=1}^N \int_{V_j} \big(\phi_\Sigma(x - \mu_j) - \phi_{\sigma_n}(x - \mu_j)\big) \, dP(\mu, \Sigma) + \sum_{j=1}^N \phi_{\sigma_n}(x - \mu_j)\big[P(V_j) - p_j\big].
\end{aligned}
$$

Note that $P(V_0) = 1 - \sum_{j=1}^N P(V_j) \leq \sum_{j=1}^N |P(V_j) - p_j|$. Hence, by triangle inequality,

$$
\begin{aligned}
\|f_P - \phi_{\sigma_n} \star F_{\sigma_n}\|_1 &\leq \sum_{j=1}^N \int_{V_j} \big\|\phi_\Sigma(\cdot - \mu) - \phi_\Sigma(\cdot - \mu_j)\big\|_1 \, dP(\mu, \Sigma) \\
&\quad + \sum_{j=1}^N \int_{V_j} \|\phi_\Sigma - \phi_{\sigma_n}\|_1 \, dP(\mu, \Sigma) + 2\sum_{j=1}^N \big|P(V_j) - p_j\big|.
\end{aligned}
$$

Recall now that $V_j = U_j \times S_{\sigma_n}$ for $j = 1, \ldots, N$ so that, by construction, $\|\mu - \mu'\| \leq \sigma_n \tilde{\varepsilon}_n^{2b_1}$ for any $\mu, \mu' \in U_j$. Also any $\Sigma \in S_{\sigma_n}$ satisfies $\det(\Sigma^{-1}) \geq 2^{-d} \sigma_n^{-2d}$, $y^T \Sigma^{-1} y \leq \|y\| \sigma_n^{-2}$ for any $y \in \mathbb{R}^d$ and $|\operatorname{tr}(\sigma_n^2 \Sigma^{-1}) - d - \log \det(\sigma_n^2 \Sigma^{-1})| \leq d \sigma_n^{2\beta}$. Hence, by using arguments similar to the proof of Lemma 2,

$$\left\| \phi_\Sigma(\cdot - \mu) - \phi_\Sigma(\cdot - \mu_j) \right\|_1 \leq \|\mu - \mu_j\| \lambda_1(\Sigma^{-1})^{1/2} \leq \sigma_n \tilde{\varepsilon}_n^{2b_1} \sigma_n^{-1} = \tilde{\varepsilon}_n^{2b_1},$$

$$\left\| \phi_\Sigma - \phi_{\sigma_n} \right\|_1 \leq \left| \operatorname{tr}(\sigma_n^{-2} \Sigma^{-1}) - d - \log \det(\sigma_n \Sigma^{-1}) \right| \leq d \sigma_n^{2\beta}$$

for any $(\mu, \Sigma) \in V_j$ and any $j = 1, \ldots, N$. Hence, for $P \in \mathcal{P}_{\sigma_n}$,

$$\| f_P - \phi_{\sigma_n} \star F_{\sigma_n} \|_1 \leq \tilde{\varepsilon}_n^{2b_1} + d \sigma_n^{2\beta} + 4 \tilde{\varepsilon}_n^{2db_1} e^{-C_3 \sigma_n^{-\kappa^*}} \lesssim \sigma_n^{2\beta}$$

since $\tilde{\varepsilon}_n^{2b_1} \leq \sigma_n^{2\beta}$ by the assumption on $b_1$. It follows that $d(f_P, \phi_{\sigma_n} \star F_{\sigma_n}) \lesssim \sigma_n^\beta$ as desired.

The last step is to control the ratio $f_P/f_0$ for any $P \in \mathcal{P}_{\sigma_n}$ in view of an application of Lemma B2 of [27]. Since $(\mu_1, \ldots, \mu_N)$ forms a $\sigma_n a_{\sigma_n}$-net of $\{x \in \mathbb{R}^d : \|x\| \leq a_{\sigma_n}\}$, the set $\{\mu \in \mathbb{R}^d : \|\mu - x\| \leq \sigma_n(a_{\sigma_n} + \tilde{\varepsilon}_n^{2b_1}/2)\}$ contains at least one $U_j$ for $j = 1, \ldots, N$. Call $J(x)$ such index $j$ and recall that, for $P \in \mathcal{P}_\sigma$,

$$P(V_{J(x)}) \geq p_{J(x)} - 2 \tilde{\varepsilon}_n^{2db_1} e^{-C_3 \sigma_n^{-\kappa^*}} \geq \tilde{\varepsilon}_n^{4db_1} - 2 \tilde{\varepsilon}_n^{2db_1} e^{-C_3 \sigma_n^{-\kappa^*}} \gtrsim \tilde{\varepsilon}_n^{4db_1}.$$

Recall that $f_0$ is bounded. For $\|x\| \leq a_{\sigma_n}$ we have

$$\frac{f_P(x)}{f_0(x)} \geq K_1 \int_{S_{\sigma_n}} \int_{\|\mu - x\| \leq \sigma_n(a_{\sigma_n}^{\tau/2} + \tilde{\varepsilon}_n^{2b_1}/2)} \phi_\Sigma(x - z) \, dP(z, \Sigma)$$

$$\geq K_2 \sigma_n^{-d} \exp\left\{ -\frac{1}{2\sigma_n^2} \left[ \sigma_n \left( a_{\sigma_n}^{\tau/2} + \tilde{\varepsilon}_n^{2b_1}/2 \right) \right]^2 \right\} P(V_{J(x)})$$

$$\geq K_2 \sigma_n^{-d} e^{-2a_{\sigma_n}^\tau} P(V_{J(x)}) \geq K_3 \sigma_n^{-d} e^{-2a_{\sigma_n}^\tau} \tilde{\varepsilon}_n^{4db_1}$$

for some constants $K_1, K_2, K_3$. Also, for every $\|x\| > a_{\sigma_n}$,

$$\frac{f_P(x)}{f_0(x)} \geq K_1 \int_{S_{\sigma_n}} \int_{\|\mu\| \leq a_{\sigma_n}} \phi_\Sigma(x - \mu) \, dP(\mu, \Sigma)$$

$$\geq K_2 \sigma_n^{-d} \int_{S_{\sigma_n}} \int_{\|\mu\| \leq a_{\sigma_n}} \exp\left\{ -\frac{1}{2\sigma^2} \|x - \mu\|^2 \right\} dP(\mu, \sigma)$$

$$\geq K_2 \sigma_n^{-d} \exp\left\{ -2\|x\|^2/\sigma_n^2 \right\} P\left( \{ \mu \in \mathbb{R} : \|\mu\| \leq a_{\sigma_n} \} \times S_{\sigma_n} \right)$$

$$\geq K_2 \sigma_n^{-d} \exp\left\{ -2\|x\|^2/\sigma_n^2 \right\} \sum_{j=1}^N P(V_j) \geq K_4 \sigma_n^{-d} \exp\left\{ -2\|x\|^2/\sigma_n^2 \right\}$$

for some constant $K_4$ because $\|x - \mu\|^2 \leq 2\|x\|^2 + 2\|\mu\|^2 \leq 4\|x\|^2$ (since $\|\mu\| \leq a_{\sigma_n} < \|x\|$) and

$$\sum_{j=1}^{N} P(V_j) = 1 - \sum_{j>N}^{M} P(V_j) = 1 - \sum_{j>N}^{M} |P(V_j) - p_j|$$

$$\geq 1 - \sum_{j=1}^{M} |P(V_j) - p_j| \geq 1 - 2\tilde{\varepsilon}_n^{2db_1} e^{-C_3\sigma_n^{-\kappa^*}}$$

and the last inequality follows from the definition of $\mathcal{P}_\sigma$. Set $\lambda = K_3\sigma_n^{-d}e^{-2a_{\sigma_n}^\tau}\tilde{\varepsilon}_n^{4db_1}$ and notice that

$$\log(1/\lambda) = \log\left(1/\tilde{\varepsilon}_n^{4db_1}\sigma_n^{-d}\right) + \log\left(\exp\{2a_0^2\log(1/\sigma_n)\}\right)$$

$$= \log\left(1/\tilde{\varepsilon}_n^{d(4b_1 - 1/\beta)}\{\log(1/\tilde{\varepsilon}_n)\}^{d/\beta}\right) + 2a_0^2\log(1/\sigma_n) \lesssim \log(1/\tilde{\varepsilon}_n)$$

since $4b_1 > 1/\beta$. Moreover, for any $P \in \mathcal{P}_\sigma$, $\{x \in \mathbb{R} : f_P(x)/f_0(x) < \lambda\} \subset \{x \in \mathbb{R} : \|x\| > a_{\sigma_n}\}$ so that

$$F_0\left\{\left(\log\frac{f_0}{f_P}\right)^2 \mathbb{1}\left(\frac{f_P}{f_0} < \lambda\right)\right\}$$

$$\leq \int_{\|x\|>a_{\sigma_n}} \left(\log\frac{f_P(x)}{f_0(x)}\right)^2 f_0(x)\,dx$$

$$\leq \int_{\|x\|>a_{\sigma_n}} \left(\log\left[\frac{\sigma_n^d}{K_4}\exp\{-2\|x\|^2/\sigma_n^2\}\right]\right)^2 f_0(x)\,dx$$

$$\leq \frac{K_5}{\sigma_n^4}\int_{\|x\|>a_{\sigma_n}} \|x\|^4 f_0(x)\,dx = \frac{K_5}{\sigma_n^4}\int_{\|x\|>a_{\sigma_n}} \|x\|^4 f_0(x)^{1/2} f_0(x)^{1/2}\,dx$$

$$\leq \frac{K_5}{\sigma_n^4}\left\{\int_{\|x\|>a_{\sigma_n}} \|x\|^8 f_0(x)\,dx \int_{\|x\|>a_{\sigma_n}} f_0(x)\,dx\right\}^{1/2}$$

$$\leq \frac{K_5}{\sigma_n^4}\left(F_0\|X\|^8\right)^{1/2}\left(F_0\{x : \|x\| > a_{\sigma_n}\}\right)^{1/2} \leq \frac{K_6}{\sigma_n^4}\left(F_0\{x : \|x\| > a_{\sigma_n}\}\right)^{1/2}$$

$$\leq \frac{K_6}{\sigma_n^4}F_0\left(E_{\sigma_n}^c\right)^{1/2} \leq \frac{K_6}{\sigma_n^4}B_0\sigma_n^{2\beta+\varepsilon+4} \leq K_7\sigma_n^{2\beta+\varepsilon}$$

for some constants $K_5, K_6, K_7$. The forth inequality follows from Cauchy–Schwarz, the sixth inequality follows from $F_0\|X\|^m < \infty$ for all $m > 0$ because of the tail condition (21) on $f_0$, the seventh and eighth inequalities follows from Proposition 1 of [27]. Since $\log x \leq (\log x)^2$ for $x > e^1$ and $\lambda < e^{-1}$ for $n$ sufficiently large, we also have that $F_0\{\log(f_0/f_P)\mathbb{1}(f_P/f_0 < \lambda)\} \leq K_8\sigma_n^{2\beta+\varepsilon}$. Now apply Lemma B2 of [27] to conclude that both $F_0\{\log f_0/f_P\}$ and

$F_0\{(\log f_0/f_P)^2\}$ are bounded by

$$K_9\big\{\log(1/\lambda)\big\}^2 \sigma_n^{2\beta} \leq A\sigma_n^{2\beta}\big\{\log(1/\tilde{\varepsilon}_n)\big\}^2 = A\tilde{\varepsilon}_n^2$$

for some positive constant $A$. Together with (32), this completes the proof.  □

## Acknowledgements

## References

[1] Barron, A., Schervish, M.J. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. *Ann. Statist.* **27** 536–561. MR1714718

[2] Bhattacharya, A. and Dunson, D.B. (2011). Sparse Bayesian infinite factor models. *Biometrika* **98** 291–306. MR2806429

[3] Carvalho, C.M., Chang, J., Lucas, J.E., Nevins, J.R., Wang, Q. and West, M. (2008). High-dimensional sparse factor modeling: Applications in gene expression genomics. *J. Amer. Statist. Assoc.* **103** 1438–1456. MR2655722

[4] Chen, M., Silva, J., Paisley, J., Wang, C., Dunson, D. and Carin, L. (2010). Compressive sensing on manifolds using a nonparametric mixture of factor analyzers: Algorithm and performance bounds. *IEEE Trans. Signal Process.* **58** 6140–6155. MR2790088

[5] Cron, A. and West, M. (2016). Models of random sparse eigenmatrices and Bayesian analysis of multivariate structure. In *Statistical Analysis for High-Dimensional Data*: *The Abel Symposium* 2014 (A. Frigessi, P. Bühlmann, K.I. Glad, M. Langaas, S. Richardson and M. Vannucci, eds.) 125–153. Cham: Springer International Publishing.

[6] Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation. The $L_1$ View. Wiley Series in Probability and Mathematical Statistics*: *Tracts on Probability and Statistics*. New York: Wiley. MR0780746

[7] De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prunster, I. and Ruggiero, M. (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.* **37** 212–229.

[8] Edelman, A. and Sutton, B.D. (2005). Tails of condition number distributions. *SIAM J. Matrix Anal. Appl.* **27** 547–560. MR2179688

[9] Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *J. Amer. Statist. Assoc.* **90** 577–588. MR1340510

[10] Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Ann. Statist.* **1** 209–230. MR0350949

[11] Ghosal, S., Ghosh, J.K. and Ramamoorthi, R.V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Statist.* **27** 143–158. MR1701105

[12] Ghosal, S., Ghosh, J.K. and van der Vaart, A.W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531. MR1790007

[13] Ghosal, S. and van der Vaart, A. (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Statist.* **35** 697–723. MR2336864

[14] Ghosal, S. and van der Vaart, A.W. (2001). Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.* **29** 1233–1263. MR1873329

[15] Gnedin, A. and Pitman, J. (2005). Exchangeable Gibbs partitions and Stirling triangles. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov.* (*POMI*) **325** 83–102, 244–245. MR2160320

[16] Goldenshluger, A. and Lepski, O. (2014). On adaptive minimax density estimation on $R^d$. *Probab. Theory Related Fields* **159** 479–543. MR3230001

[17] Gorur, D. and Rasmussen, C. (2009). Nonparametric mixtures of factor analyzers. In *Signal Processing and Communications Applications Conference*, 2009. *SIU* 2009. *IEEE* 17*th* 708–711. Cambridge, MA: MIT Press.

[18] Kruijer, W., Rousseau, J. and van der Vaart, A. (2010). Adaptive Bayesian density estimation with location-scale mixtures. *Electron. J. Stat.* **4** 1225–1257. MR2735885

[19] Lijoi, A., Mena, R.H. and Prünster, I. (2005). Hierarchical mixture modeling with normalized inverse-Gaussian priors. *J. Amer. Statist. Assoc.* **100** 1278–1291. MR2236441

[20] Lijoi, A., Prünster, I. and Walker, S.G. (2005). On consistency of nonparametric normal mixtures for Bayesian density estimation. *J. Amer. Statist. Assoc.* **100** 1292–1296. MR2236442

[21] Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. I. Density estimates. *Ann. Statist.* **12** 351–357. MR0733519

[22] MacEachern, S.N. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *J. Comput. Graph. Statist.* **7** 223–238.

[23] Matthaiou, M., McKay, M.R., Smith, P.J. and Nossek, J.A. (2010). On the condition number distribution of complex Wishart matrices. *IEEE Trans. Commun.* **58** 1705–1717.

[24] Muirhead, R.J. (1982). *Aspects of Multivariate Statistical Theory*. New York: Wiley. MR0652932

[25] Müller, P., Erkanli, A. and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83** 67–79. MR1399156

[26] Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Related Fields* **92** 21–39. MR1156448

[27] Shen, W., Tokdar, S.T. and Ghosal, S. (2013). Adaptive Bayesian multivariate density estimation with Dirichlet mixtures. *Biometrika* **100** 623–640. MR3094441

[28] Tokdar, S.T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā* **68** 90–110. MR2301566

[29] Walker, S.G., Lijoi, A. and Prünster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. *Ann. Statist.* **35** 738–746. MR2336866

[30] West, M. (2003). Bayesian factor regression models in the "large *p*, small *n*" paradigm. In *Bayesian Statistics*, 7 (*Tenerife*, 2002) 733–742. Oxford Univ. Press, New York. MR2003537

[31] Wu, Y. and Ghosal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electron. J. Stat.* **2** 298–331. MR2399197

[32] Wu, Y. and Ghosal, S. (2010). The $L_1$-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. *J. Multivariate Anal.* **101** 2411–2419. MR2719871